



# INTERAÇÕES ENTRE *BIG DATA* E *CLOUD COMPUTING*

**© 2019 por Editora e Distribuidora Educacional S.A.**

Todos os direitos reservados. Nenhuma parte desta publicação poderá ser reproduzida ou transmitida de qualquer modo ou por qualquer outro meio, eletrônico ou mecânico, incluindo fotocópia, gravação ou qualquer outro tipo de sistema de armazenamento e transmissão de informação, sem prévia autorização, por escrito, da Editora e Distribuidora Educacional S.A.

**Presidente**

Rodrigo Galindo

**Vice-Presidente de Pós-Graduação e Educação Continuada**

Paulo de Tarso Pires de Moraes

**Conselho Acadêmico**

Carlos Roberto Pagani Junior

Camila Braga de Oliveira Higa

Carolina Yaly

Giani Vendramel de Oliveira

Juliana Caramigo Gennarini

Nirse Ruscheinsky Breternitz

Priscila Pereira Silva

Tayra Carolina Nascimento Aleixo

**Coordenador**

Nirse Ruscheinsky Breternitz

**Revisor**

Andre Filipe de Moraes Batista

**Editorial**

Alessandra Cristina Fahl

Beatriz Meloni Montefusco

Daniella Fernandes Haruze Manta

Hâmila Samai Franco dos Santos

Mariana de Campos Barroso

Paola Andressa Machado Leal

Dados Internacionais de Catalogação na Publicação (CIP)

---

L864i Lopes, Aimar Martins  
Interações entre Big Data e Cloud Computing/ Aimar  
Martins Lopes, – Londrina: Editora e Distribuidora  
Educacional S.A. 2019.  
141 p.

ISBN 978-85-522-1572-1

1. Banco de dados. 2. Computação em nuvem I. Lopes,  
Aimar Martins. II. Título.

CDD 004

---

Thamiris Mantovani CRB: 8/9491

2019

Editora e Distribuidora Educacional S.A.  
Avenida Paris, 675 – Parque Residencial João Piza  
CEP: 86041-100 — Londrina — PR  
e-mail: editora.educacional@kroton.com.br  
Homepage: <http://www.kroton.com.br/>



---

## SUMÁRIO

Apresentação da disciplina	04
<i>Big Data</i> : fundamentos, infraestrutura e interfaces	05
Gerenciamento de <i>Big Data</i> : coleta e processamento de dados	24
Aplicação da ciência de dados no gerenciamento empresarial	42
<i>Big Data Analytics</i>	66
Algoritmos de aprendizado de máquina para minerar os dados	88
<i>Cloud computing</i> e <i>Big Data</i>	108
Estruturas de programação em nuvem	126

## Apresentação da disciplina

Interações entre Big Data e *cloud computing* são conceitos fundamentais nas áreas da tecnologia da informação e administração de negócios. A primeira justifica-se pelo conjunto de tecnologias envolvidas e o motivo pelo qual as organizações, em virtude da internet, modificam a forma de coletar e analisar dados, e com eles toma decisões melhores e cria novos modelos de negócios. A segunda, justifica-se pela modificação da abordagem na gestão dos recursos computacionais (processamento, armazenamento, servidores, etc.) e formas de processamento e armazenamento dos dados.

Estudaremos os fundamentos de Big Data, arquitetura Hadoop para coleta, processamento e armazenamento de dados, veremos a importância da ciência de dados, como dados podem mudar os modelos de negócios e, finalmente, os algoritmos, o que são, quais são os de uso mais comum em Big Data e os conceitos de *machine learning* e *deep learning*.

Com relação à estrutura de *cloud computing*, veremos como é seu funcionamento, os modelos de serviços e implantação de nuvem, como diferenciar e analisar um cloud, e por que a relação da computação em nuvem com o Big Data é importante no conceito de análise de dados. Contudo, há algumas dificuldades, e neste item conheceremos os desafios que Big Data enfrenta na computação em nuvem. Veremos como analisar os modelos de serviços em nuvem, conheceremos o MS Azure e o Google Cloud Platform e algumas linguagens de programação usadas *in cloud*.



# ***Big Data: fundamentos, infraestrutura e interfaces***

Autor: Nome do autor da disciplina

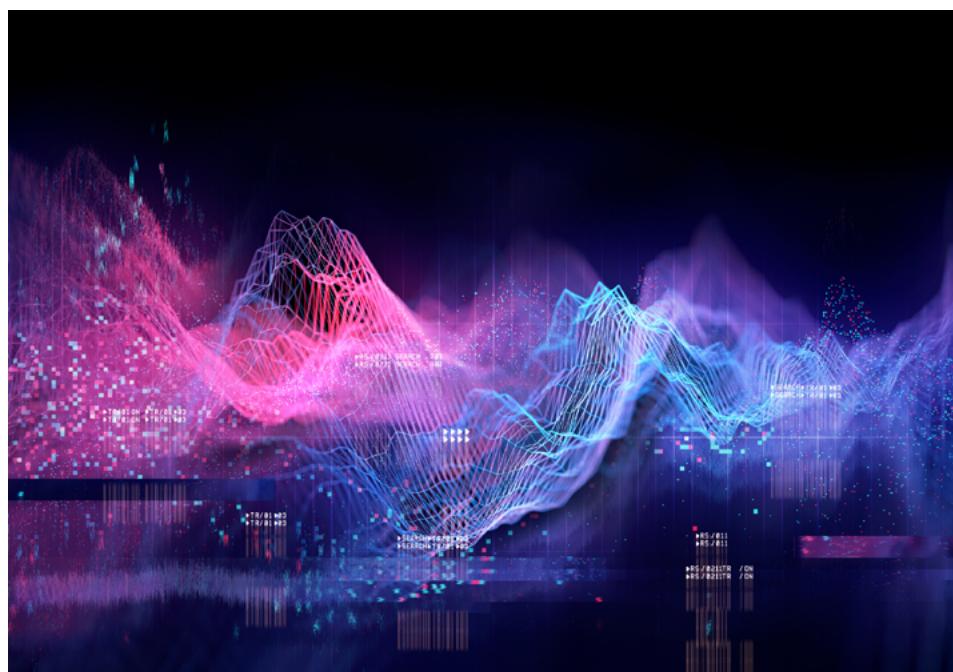
## **► Objetivos**

- Compreender os fundamentos do Big Data.
- Conhecer a estrutura 5 Vs do Big Data.
- Descrever as interfaces de potencial de uso com Big Data.

## 1. Big Data

O termo Big Data é referenciado em qualquer área, seja ela ciência, indústria, negócios, cultura, saúde, etc., pois está relacionado a captura e análise de muitos dados, tendo como característica principal o volume, a variedade e a velocidade, tanto de criação como movimentação, tendo como consequência a força de influenciar e modificar a sociedade de maneira drástica. A Figura 1 mostra um gráfico futurístico ilustrando uma análise de dados complexa, em que as cores representam informações diferentes e os pontos de intensidade.

Figura 1 – Gráfico futurístico de análise de dados



Fonte: solarseven / iStock

Qual o conceito envolvido com o Big Data? A sua evolução está relacionada com a capacidade da humanidade em analisar dados, evolução da computação no sentido de processamento e análise de dados, a comunicação com a internet e suas aplicações que geram bilhões de dados por segundos, estatística, arquitetura de software complexa e os novos modelos de negócios que utilizam os dados avidamente.

Como o termo é relativo, abrangente e de vários entendimentos, seguem algumas definições.

Big Data significa um grande volume de informações de alta variedade e velocidade que exige formas inovadoras e econômicas de análise para melhor tomada de decisões e automação de processos (GARTNER, 2019).

Segundo a SAS (2019):

Big Data é um termo que descreve o grande volume de dados — tanto estruturados quanto não estruturados — que sobrecarrega as empresas diariamente. Mas não é a quantidade de dados disponíveis que importa; é o que as organizações fazem com eles. O Big Data pode ser analisado para obter *insights* que levam a decisões melhores e ações estratégicas de negócio.

Sem dúvida, a ação organizacional sobre o que fazer com os dados, afirmado pela SAS, e os *insights* são o que faz a diferença nas estratégias.

Apesar do termo existir há algum tempo, ainda há muita confusão relativa ao seu significado. O conceito sempre está evoluindo e se modificando, pois é uma das grandes forças motrizes da transformação digital. Portanto, Big Data está relacionado com coletar dados e a capacidade de uso para obter vantagem em diversas áreas, incluindo negócios (MARR, 2019a).

Para Steve Perry, Big Data está relacionado com o significado do dado, é um processo que está cada vez mais acelerado e com mais fontes e formatos variados de dados. Afirma que, em breve, chamaremos isso de Big Meaning (grande significado), pois o que realmente importa é o valor (significado) dos dados e não a sua quantidade (PERRY, 2019).

## PARA SABER MAIS



Uma tecnologia só é útil se resolver algum problema. Há muitos dados históricos, mas novos dados são gerados

por aplicativos de rede social, cliques em sites, fluxo de aplicativos da web, dispositivos, sensor IoT entre outros. A quantidade de dados gerada é enorme e continua crescendo em muitos formatos diferentes. O valor dos dados tem significado quando podemos extrair algo deles, e obter valor desses dados não é tarefa fácil (PERRY, 2017).

## ► 2. Uma breve história

Em vez de ser uma única tecnologia, o Big Data é um ecossistema de técnicas e tecnologias coordenadas que extraem valor comercial das montanhas de dados produzidos no mundo atual. A relatividade da definição se dá pela palavra *big*: questionamos o que é ser grande. Isso depende. Se um laboratório de análises clínicas reunir seus dados de um ano, pode ser que, para o estudo dos serviços executados, isso seja grande.

John Graunt, em 1663, reuniu uma série de dados para estatisticamente estudar a peste bubônica na Europa e, talvez para ele, os dados que possuía tinham o sentido de grande (FOOTE, 2017). Uma concepção mais moderna envolve o desenvolvimento de computadores, smartphones, internet IoT, rede social e o tamanho das organizações.

Os fundamentos do Big Data tiveram início com a solução de um problema da U. S. Census Bureau em 1880, quando Herman Hollerith criou uma máquina de tabulação que reduziu o tempo que seria de dez anos para processar o censo em três meses. Em 1927, o engenheiro Fritz Pleumer desenvolveu a fita magnética, possibilitando armazenar dados de forma mais eficiente. Durante a Segunda Guerra Mundial, a Inglaterra criou a máquina Colossus, que escaneava 5 mil caracteres por segundo e possibilitou a interpretação dos códigos secretos de guerra

da Alemanha. Em 1945, John Von Neumann publicou o artigo *Electronic Discrete Variable Automatic Computer (EDVAC)*, sobre o armazenamento de programas e a arquitetura de computadores, que se mantém até hoje (FOOTE, 2017).

Mais recentemente, inclui-se na lista a criação da internet, com o nome inicial de ARPANET, em 1969, nos Estados Unidos, para conectar computadores. Com sua evolução, atualmente, toda a sociedade mundial trafega dados pela rede. Em 1989, Tim Berners-Lee criou o conceito Word Wide Web (WWW) (BARROS, 2013). Isso possibilitou o acesso a vários endereços da rede de forma rápida e uma esplêndida evolução de fluxo de dados diversos pela internet, seja ele texto, áudio, vídeo e foto. Mas, principalmente, possibilitou o compartilhamento de informação na internet. Conclusão: a criação e circulação de dados pelo mundo aumentou significativamente. Os conceitos envolvidos na WWW são: HTML (HyperText Markep Language), URL (Uniform Resource Locator) e HTTP (HyperText Transfer Protocol). A Figura 2 ilustra uma página WWW com o endereço URL.

Figura 2 – Página WWW



Fonte: crstrbrt / iStock

A indústria da computação pessoal também tem sua contribuição. Os microcomputadores ocuparam intensamente os espaços no mercado, especialmente com a Apple e Microsoft, por volta de 1977. Sua evolução também acompanha a evolução e disseminação da internet e, consequentemente, do Big Data. Os preços dos microcomputadores caíram muito nos anos de 1980 e 1990, facilitando seu uso por grande parte da população ao redor do mundo, pois, com a evolução da comunicação, a internet chegou também ao indivíduo comum.

O cenário está montado para a evolução do Big Data, temos computadores, internet (comunicação), novos softwares, redes e softwares que conectam pessoas ao redor do mundo. Diante disso, surgem as redes sociais, precursoras de grande volume de dados.

Contudo, foi em 1993 que a CERN (Organização Europeia para a Investigação Nuclear), local onde Tim Berners-Lee era consultor, promoveu o compartilhamento das informações dos pesquisadores por meio da WWW, divulgando a ideia e deixando-a à disposição para que qualquer pessoa pudesse usar e desenvolver aplicações, com isso a internet se proliferou e ficou à disposição para qualquer pessoa usar e desenvolver aplicações (WENKEL, 2016). Para Foote (2013), esse foi um fator-chave para a evolução da WEB como um todo, pois possibilitou que pessoas do mundo inteiro pudessem ter acesso e que organizações pudessem prover conexões de internet para todos e para tudo.

Somente na virada do século 21 é que a web explodiu com o surgimento de várias organizações “.com” e diversos modelos de negócios: os hoje conhecidos como *e-commerce*s. Esse era o cenário com muito combustível à disposição para a geração de dados e sua movimentação em toda rede mundial. Com isso, o termo Big Data passa a ficar mais concreto.

Porém, mais elementos foram surgindo, a IoT (Internet of Things) se fortaleceu por volta de 2013, com o uso de várias tecnologias, tais como: internet, sistemas microeletrônicos e mecânicos, programação embarcada, comunicação wireless, GPS, etc. Todos esses elementos geram ou transmitem dados das pessoas, casas, organizações e de todas as coisas, conclui-se então que mais dados entraram em circulação por todo o mundo.

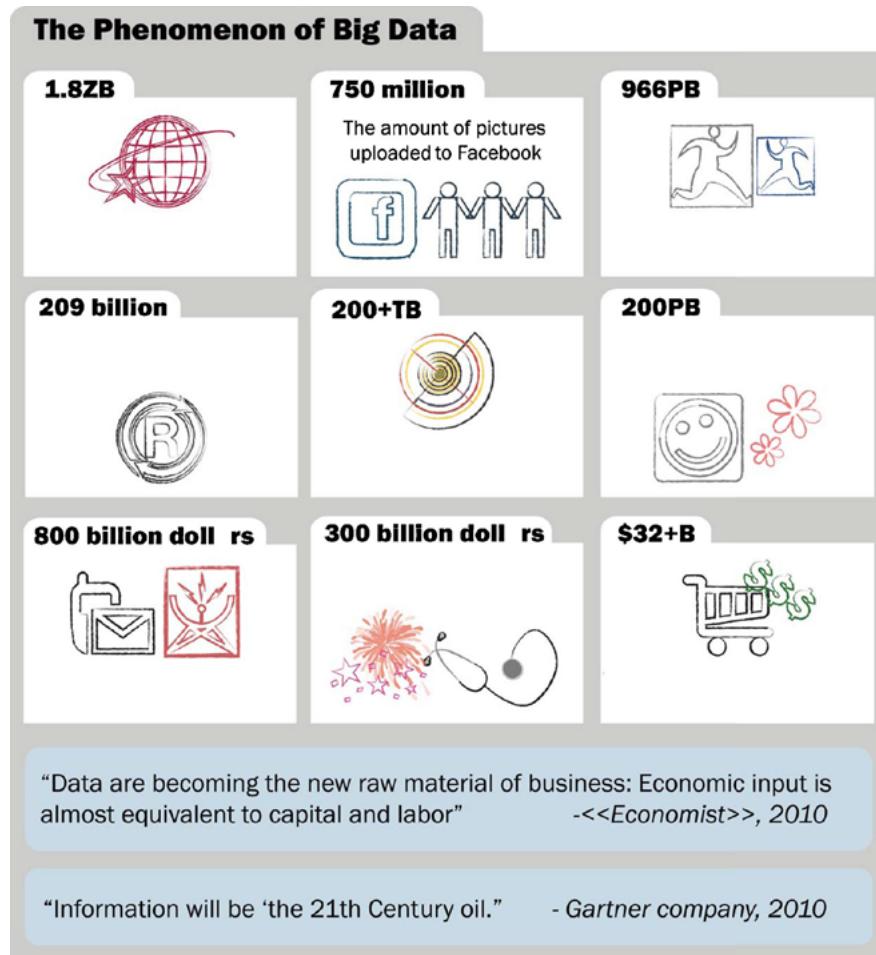
Por volta de 2003, a Google já utilizava sistema de armazenamento de dados distribuído e também processava dados distribuído e paralelo na rede, cujos nomes das tecnologias são GFS (Google File System) e MapReduce. Em seguida, por volta de 2005, o sistema Hadoop foi criado como software aberto para processar e armazenar de forma distribuída dados que circulam pela internet. Este também foi um marco para o Big Data, pois com o volume que circulava na internet, não havia sistema capaz de estruturar e analisar.

Nesse cenário, os hardwares continuam tendo grande importância, pois são os executores das tarefas. Tanto a evolução dos processadores como os mecanismos para armazenar informações, e claro os de comunicação, formam o motor do fluxo de dados.

A computação em nuvem (*cloud computing*) utiliza toda essa engrenagem que possibilita o compartilhamento de recursos pela internet. Com ela, após o ano de 1990, temos início da oferta de tecnologia como serviço, são elas IaaS, SaaS e PaaS. Tudo em larga escala.

No Brasil, aproximadamente no ano de 2010, começamos a usufruir dessa avalanche de tecnologias com mais facilidade, e não somente no Brasil, mas no mundo, verificamos mudanças de comportamento humano, formas de trabalho, redes sociais, modelos de negócios e muito mais. Sendo que no que se refere ao Big Data, seu uso é diverso. A Figura 3 ilustra o fenômeno Big Data.

Figura 3 – Fenômeno Big Data



Fonte: CHEN, 2014.

### ► 3. Estrutura do Big Data

Big Data leva a imaginar um monte de dados não sistematizados que passam por vários processamentos até fornecerem algum tipo de informação. Para entendermos a sua estrutura, consideramos a origem e o formato. Podemos afirmar que as fontes de onde provêm os dados são as mais diversas possíveis, simplificaremos, em dispositivos diversos, dados estruturados de banco de dados organizacionais e/ ou não estruturados que circulam pelas redes sociais e muitos outros. Quanto ao armazenamento, também são utilizados vários métodos (PROVOST, 2013).

Dentre as características que tratam Big Data, muitas têm surgido e forçam mudanças no conceito, por isso apresentaremos as características que Bernard Marr cita, trata-se de um especialista que acompanha a evolução da tecnologia. Inicialmente, o Big Data era referenciado por 3 Vs: volume, velocidade e variedade. Esse conceito tem se modificado, alguns definem até 8 ou mais Vs, mas trataremos somente de 5 Vs, por serem suficientes. Para Bernard, os 5 Vs são: volume, velocidade, variedade, veracidade e valor (MARR, 2014).

Para maior entendimento, seguem seus significados na Figura 4 abaixo.

Figura 4 – Significado dos 5 Vs



Fonte: elaborado pelo autor.

**Volume** – Grande volume de dados gerados a todo momento por fontes diversas espalhadas pelo mundo, tais como: e-mails, mensagens de redes sociais, imagens, vídeos, dados de sensores, dados gerados por navegação pela web, etc. A quantidade de dados é surreal. O Facebook sozinho transaciona mais de 10 bilhões de mensagens por dia (MARR, 2014).

**Velocidade** – Refere-se à velocidade de geração dos dados, relação de velocidade, transporte entre os sistemas, conexão com a internet e volume. Quando mais rápido uma organização acessa o dado, mais assertiva pode ser sua decisão e predição.

**Variedade** – Vários tipos de dados existem espalhados pela internet que estão relacionados com os sistemas de informação e dispositivos, fontes diferentes, e diferentes são os métodos de transporte, tratamento e armazenamento. Contudo, em um determinado momento, é necessário conhecer o tipo mais adequado para utilizar as ferramentas, os algoritmos e modelos de análise mais adequados. Cerca de 80% dos dados disponíveis no mundo são desestruturados, o que dificulta seu condicionamento e sua tabulação para que seja feito um melhor tratamento e aquisição de informações a partir deles (MARR, 2014).

**Veracidade** – O dado bom dá confiança, portanto, a confiabilidade dos dados é fundamental dentro desse universo caracterizado por volume, formato, velocidade e variedade. O desafio causado pela grande variedade e pelo formato dos dados é a dificuldade de tratá-los completamente e obter as informações mais precisas e conclusivas (MARR, 2014).

**Valor** – Refere-se ao significado que o dado pode dar para atender a uma necessidade ou resolver um problema. O dado que não representa um valor não deve receber atenção. Para Marr (2014), o valor resulta do acesso aos dados e ao conhecimento obtido da sua análise para estruturar de maneira definida e objetiva a proposta de ação

empresarial e tomar as melhores decisões organizacionais. A Figura 5 ilustra um homem de negócios interpretando gráficos.

Figura 5 – Homem de negócios interpretando dados



Fonte: NicoElNino / istock

## ► 4. Interface e possibilidade

Há uma gama enorme de possibilidades de uso relacionado ao Big Data. Estamos ainda vendo somente a ponta do iceberg, muita coisa está para mudar e ser criada.

### ASSIMILE

O Big Data tem o princípio de que quanto mais se sabe de algo, mais confiança se tem para obter novos *insights* e fazer previsões sobre o futuro. Quando se comparam dados, relacionamentos ocultos começam a surgir, e eles permitem aprender a tomar decisões inteligentes. O processo envolve

criação de modelos com base nos dados, execução de simulações, aprimoramento do valor e monitoramento dos resultados (MARR, 2019b).

No artigo, *Big Data in Practice*, Bernard Marr (2019b) afirma que Big Data é algo muito profundo e cita dez áreas em que é possível ter excelente vantagens com seu uso, vamos a elas:

## 1. Entendendo e direcionando o cliente

É a área que mais utiliza Big Data, é usada para compreender o comportamento e a preferência do cliente. A Figura 6 ilustra a intensidade do uso dos aplicativos em redes sociais.

Figura 6 – Significado dos 5 Vs



Fonte: alexsl / istock (1094102054).

## 2. Entendendo e otimizando os processos de negócios

O comércio pode melhorar seus estoques com base na predição de dados da rede social, tendência de pesquisa e previsão de tempo. Isso reduziria o custo envolvido em gestão de estoque, despesa de compra, recebimento, movimentação e armazenamento ao longo do tempo, pois aprimora o sistema de *just in time*.

## 3. Qualificação pessoal e otimização de desempenho

Big Data não é somente para uso de organizações e governos, o indivíduo pode se beneficiar pela geração de dados de dispositivos vestíveis, tais como os relógios e braceletes. Esses dispositivos coletam dados de seu corpo a todo momento, podem coletar nível de glicemia, frequência cardíaca, etc.

## 4. Melhorar a saúde individual e a saúde pública

A análise de dados habilita a decodificação de DNA em minutos e possibilita encontrar nova cura, melhoria de tratamento e prever padrão de doenças.

## 5. Melhora no desempenho esportivo

Muitos esportes têm adotado o Big Data para analisar vídeos, equipamentos esportivos, rastreio de atletas para acompanhar o sono e a alimentação, bem como o comportamento e estado emocional que o atleta apresenta nas redes sociais.

## 6 – Melhorando a ciência e a pesquisa

O CERN, laboratório de física nuclear, possui o maior e mais poderoso acelerador de partículas do mundo, o Hadron Collider. Ele é capaz de gerar em seus experimentos com o universo uma enorme quantidade de dados que são analisados por um poderoso centro

de computação. Dados governamentais podem ser acessados por pesquisadores que criam novos cenários para a ciência.

## 7. Otimizando máquinas e desempenho de dispositivos

O Big Data auxilia na inteligência e autonomia das máquinas. Por exemplo, o carro autônomo equipado com sensores, câmeras, GPS e computadores.

## 8. Melhora da segurança e aplicação da lei.

O Big Data é utilizado intensamente para melhorar a segurança, em que agências de segurança mundial detectam intenções terroristas, investigam suspeitos, previnem ataques cibernéticos e financeiros.

## 9. Melhorando e otimizando as cidades

O volume e o fluxo de dados permitem que as cidades otimizem o tráfego com base em informações em tempo real, mídia social e dados meteorológicos. Podem ser utilizados para o controle de energia, água, semáforos, etc.

## 10. Negociação financeira

Alto-Frequency Trading (HFT) é uma área com grande potencial para Big Data e está sendo muito usado atualmente. Algoritmos para manipular dados são usados para tomar decisões de negócios comerciais, exploram as informações em busca de condições personalizáveis e oportunidades de negociação.

São tantas possibilidades que as organizações e os desenvolvedores de software, os estatísticos e matemáticos se envolvem profundamente na criação de soluções para compreender e extrair *insight* desses dados.



## TEORIA EM PRÁTICA

O dados não estruturados têm um grande potencial.

Existem muitos multiplicadores de dados, incluindo humanos, máquinas e processos de negócios, e o volume de dados cresce exponencialmente. Espera-se que os dados de saúde, seguros e os dados de fabricação cresçam enormemente a cada ano, sendo que mais de 80% desses dados são desestruturados e incapazes de serem processados por soluções existentes. As informações valiosas estão escondidas em documentos, e-mails, batepapos, transcrições de centrais de atendimento, conteúdo de mídia social, comentários de clientes e relatórios de setor.

Enquanto a análise estruturada fornece o que, onde e quando de um desafio de negócios, análise de conteúdo não estruturada fornece o porquê e como. Isso ajuda empresas a antecipar e identificar defeitos de produtos, melhorar o design de produtos, o gerenciamento de recursos e serviços, reduzir a rotatividade, identificar concorrentes e otimizar os gastos com marketing (REDDY, 2018).

Forneça um exemplo e uma solução de uma situação em que dados podem ser analisados para resolver um problema.



## VERIFICAÇÃO DE LEITURA

1. Big Data está relacionado com dados e sua interpretação, sendo assim, analise as afirmações abaixo e assinale a alternativa INCORRETA.

- a. A evolução do Big Data se relaciona com a capacidade da humanidade em analisar dados e a evolução da computação.
  - b. Big Data se relaciona com processamento, análise de dados, comunicação com a internet e suas aplicações que geram bilhões de dados por segundos.
  - c. Big Data significa um grande volume de informações de alta variedade gerado com grande velocidade.
  - d. O termo Big Data já existe há muito tempo e serve para armazenar dados.
  - e. Big Data é uma vasta variedade de dados estruturados e não estruturados que diariamente invade organizações.
2. O histórico da evolução da computação, internet e outras tecnologias faz parte do surgimento do Big Data. Avalie as afirmativas a seguir e depois assinale a alternativa que contempla as afirmativas que contribuíram para o Big Data.
  - I. John Von Neumann publicou o artigo *Electronic Discrete Variable Automatic Computer (EDVAC)*, sobre o armazenamento de programas e a arquitetura de computadores em 1945.
  - II. Em 1989, Tim Berners-Lee criou o conceito Word Wide Web (WWW). Os conceitos envolvidos na WWW são: HTML (HyperText Markep Language), URL (Uniform Resource Locator) e HTTP (HyperText Transfer Propocol).
  - III. Os preços dos microcomputadores caíram muito nos anos de 1980 e 1990, isso permitiu seu uso por

grande parte de pessoas ao redor do mundo. Com a evolução da comunicação, a internet chegou também ao indivíduo comum.

IV. A IoT (Internet of Things) se fortaleceu por volta de 2013, com o uso de várias tecnologias, tais como: internet, sistemas microeletrônicos e mecânicos, programação embarcada e comunicação wireless, GPS.

- a. I e III.
- b. II e III.
- c. I, II e III.
- d. I e II.
- e. I, II, III e IV.

3. Para Provost (2013), Big Data é um vasto conjunto de dados que sofre vários processamentos até fornecer algum tipo de informação com estrutura de origem e formato.

#### POR TANTO

As fontes de onde provêm os dados são estruturadas de banco de dados organizacionais e redes sociais.

Com base nas informações dadas e na relação proposta entre elas, é CORRETO afirmar que:

- a. As duas afirmações são verdadeiras e a segunda justifica a primeira.
- b. As duas afirmações são verdadeiras e a segunda não justifica a primeira.
- c. As duas afirmações são falsas.
- d. A primeira afirmação é verdadeira e a segunda é falsa.
- e. A primeira afirmação é falsa e a segunda é verdadeira.

## Referências bibliográficas

BARROS, T. **Internet completa 44 anos:** relembre a história da web. Disponível em: <https://www.techtudo.com.br/artigos/noticia/2013/04/internet-completa-44-anos-relembre-historia-da-web.html>. Acesso em: 3 jun. 2019.

CHEN, M.; MAO, S. LIU, Y. Big Data: A Survey. 2014. **Springer Science+Business Media**, New York, p. 71-209, 22 jan. 2014.

FOOTE, K. D. **A Brief History of Big Data.** DATAVERSITY. Disponível em: <https://www.dataversity.net/brief-history-big-data/#>>. Acesso em: 1 jun. 2019.

GARTNER. **Big Data.** Disponível em: <https://www.gartner.com/it-glossary/big-data/>. Acesso em: 3 jun. 2019.

MARR, B. **Big Data:** The 5Vs Everyone Must Know. Disponível em: <https://www.linkedin.com/pulse/20140306073407-64875646-big-data-the-5-vs-everyone-must-know>. Acesso em: 6 mai. 2019.

\_\_\_\_\_. **Big Data in practice.** 2019b. Disponível em: <https://www.bernardmarr.com/default.asp?contentID=1076>. Acesso em: 3 jun. 2019.

\_\_\_\_\_. **What is Big Data.** 2019a. Disponível em: <https://www.bernardmarr.com/default.asp?contentID=766>. Acesso em: 3 jun. 2019.

PERRY, J. S. **What is Big Data? More than volume, velocity and variety...** Disponível em: <https://developer.ibm.com/dwblog/2017/what-is-big-data-insight/>. Acesso em: 3 jun. 2019.

PROVOST, F.; FAWCETT, T. **Data Science for Business:** What you need to know about Data Mining and Data-Analytic think. Disponível em: <https://www.pdfdrive.com/data-science-for-business-what-you-need-to-know-about-data-mining-and-data-analytic-thinking-d170193185.html>. Acesso em: 30 ago 2019.

REDDY T. **5 ways to turn data into insights and revenue with cognitive content analytics.** Disponível em: <https://www.ibmbigdatahub.com/blog/5-ways-turn-data-insights-and-revenue-cognitive-content-analytics>. Acesso em: 3 jun. 2019.

SAS. **Big Data.** O que é e qual a sua importância? Disponível em: [https://www.sas.com/pt\\_br/insights/big-data/what-is-big-data.html](https://www.sas.com/pt_br/insights/big-data/what-is-big-data.html). Acesso em: 3 jun. 2019.

WENKEL, R. **Há 25 anos World Wide Web era apresentada ao mundo.** Disponível em: <https://www.dw.com/pt-br/h%C3%A1-25-anos-world-wide-web-era-apresentada-ao-mundo/a-19451351>. Acesso em: 3 jun. 2019.

## **Gabarito**

### **Questão 1 – Resposta D**

**Resolução:** a afirmação está INCORRETA, pois, considerando a afirmação de Marr, o termo existe há algum tempo e ainda há muita confusão relativa ao significado. O conceito sempre está evoluindo e se modificando, pois é uma das grandes forças motrizes da transformação digital. Portanto, Big Data está relacionado com a coleta de dados e a capacidade de uso para obter vantagem em diversas áreas, incluindo negócios (MARR, 2019a).

### **Questão 2 – Resposta E**

**Resolução:** todas as afirmações estão corretas e as evoluções contribuíram para o surgimento do Big Data de uma certa forma. Consulte a Leitura Fundamental para maiores informações.

### **Questão 3 – Resposta D**

**Resolução:** a primeira afirmação é verdadeira, mas a segunda é falsa. As fontes de onde provêm os dados são as mais diversas possíveis, simplificaremos, em dispositivos diversos, dados estruturados de banco de dados organizacionais e dados não estruturados que circulam pelas redes sociais e muitos outros.



# Gerenciamento de *Big Data*: coleta e processamento de dados

Autor: Aimar Martins Lopes

## ► Objetivos

- Compreender a arquitetura Hadoop.
- Capacitar para descrever o funcionamento do sistema de arquivo HDFS e os programas MapReduce.
- Descrever como operam os componentes integrados do Hadoop.
- Conhecer e identificar a aplicabilidade de projetos para Hadoop.

## 1. Gerenciamento dos dados

O *Big Data* requer uma estrutura para gerenciamento dos dados, de um lado, temos o armazenamento tradicional com banco de dados relacional, mais indicado e adotado para análise de dados estruturada (campos conhecidos e determinados) provenientes de vários sistemas e com capacidade de produzir *insights* descritivos conhecidos; de outro lado, temos, por exemplo, a plataforma Hadoop, mais indicada para análise de dados semiestruturados e não estruturados ou quando se deseja atender a uma necessidade ou resolver um problema não específico.

Quando lidamos com dados, precisamos ter cuidado em selecionar e entender o funcionamento do local de armazenamento. Os dados armazenados nos sistemas tradicionais passam por um rigor de tratamento até chegar à base de dados. Os desenvolvedores e os consumidores de dados sabem que o fluxo é localização, processos de qualidade e limpeza, cruzamento, enriquecimento, tabulação, metadado, modelagem e, finalmente, análise e armazenamento. A próxima etapa é disponibilizá-lo para relatórios e *dashboards*.

Enquanto o dado estruturado requer uma acurácia enorme, os repositórios de *Big Data*, muitas vezes, se iniciam com pouca qualidade e controle, devido ao custo de preparo poder ser proibitivo. Alguns analistas de dados decidem o que armazenar de acordo com sua percepção de valor. Contudo, numa plataforma Hadoop, a organização pode armazenar com fidelidade as transações, os cliques, tuítes, postagens de rede social e outras coisas de forma intacta.

O dado numa plataforma Hadoop pode ter pouco valor hoje ou não ser quantificado, mas isso pode mudar no futuro, em alguns casos, ele pode significar um problema não respondido. Portanto, quanto mais um determinado dado armazenado tenha valor, mais rigoroso e limpo será o processo de transformação e armazenamento.

No entanto, outras relações podem ser consideradas para quantificar um valor para o dado, por exemplo, o retorno financeiro do investimento em Big Data, relação do valor retornado pelo dado armazenado com o custo de computação ou processamento. O custo do dado processado no sistema *warehouse* é relativamente alto, isso necessariamente não é ruim, pois os dados têm valor em virtude de serem comprovados e conhecidos, ao contrário do sistema Hadoop que tem custo de processamento baixo.

Uma plataforma de *Big Data* pode armazenar todas as transações de um negócio para depois tentar obter valor por meio de processamento, embora, às vezes, essa estratégia não seja recomendada. É interessante deixar os dados adormecidos por um tempo no sistema *Big Data* e, quando descobrir um valor comprovado, confiável e sustentado, realizar a migração para o *warehouse*.

## PARA SABER MAIS



Qual a relação das tecnologias Hadoop e Big Data? Por que ele é importante? A resposta para essa pergunta está nas próprias características do Hadoop, pois trata-se de programas de código aberto e disponível, todos podem ter acesso, usar e até modificar. Por ser flexível, organizações criam projetos específicos e aderentes à sua arquitetura, que se somam a outros projetos disponíveis. Isso o tornou popular e faz com que seja adotado cada vez mais por organizações, sendo esse mais um motivo para seu uso, além, é claro, das suas excelentes características técnicas.

## ► 2. Hadoop

A Apache Software Foundation, organização que desenvolve vários softwares de código aberto, desenvolveu em linguagem Java um sistema de arquivos distribuído em cluster para computação com dados em grande escala, chamado Hadoop. A Figura 7 mostra a imagem que identifica o Hadoop.

Figura 7 – Imagem do Hadoop



Fonte: HADOOP, 2019.

O Hadoop foi concebido seguindo os princípios do GFS (Google File System) e no processamento distribuído MapReduce. O MapReduce quebra em tarefas um trabalho, realiza processamento paralelo massivo pela rede e reduz a tarefa em novas tarefas para manipular dados e armazenar em servidores.

A arquitetura do Hadoop trabalha para produzir resultados com base na alta escalabilidade e processamento *batch* distribuído. Não é um projeto que visa operações *real time*, de velocidade de tempo de resposta ou armazenamento, foi projetado para busca e descoberta, com a intenção de fazer algo desconhecido ou impossível se tornar próximo de uma possibilidade de análise.

A metodologia Hadoop é construída em torno de um modelo de função para dados, ao contrário de dados para função, ou seja, os dados determinam um modelo para análise, isso é possível pela presença de muitos dados que são analisados por programas.

A arquitetura Hadoop, como já dissemos, é dividida em duas partes:

- HDFS – Hadoop Distributed File System;
- Modelo de programação MapReduce.

A programação MapReduce opera de forma redundante, tanto no ambiente de armazenamento de dado no cluster como na tolerância de falha no processamento, ou seja, os problemas de erro são automaticamente solucionados por vários servidores de processamento no cluster. Esse ambiente de redundância permite o dimensionamento de cargas de processamento em clusters de máquinas diversas espalhadas pelo mundo com processamento de custo baixo, consequentemente, todo esse cenário é o que possibilita os projetos de Big Data.

## 2.1 Os componentes do Hadoop

A arquitetura Hadoop é composta de cinco processos integrados responsáveis pelo trabalho, eles são agrupados da seguinte forma:

Modelo de programação MapReduce é composto pelos processos:

- NameNode;
- DataNode;
- SecondaryNameNode.

O outro conjunto de processos pertence ao HDFS:

- JobTracker;
- TaskTracker.

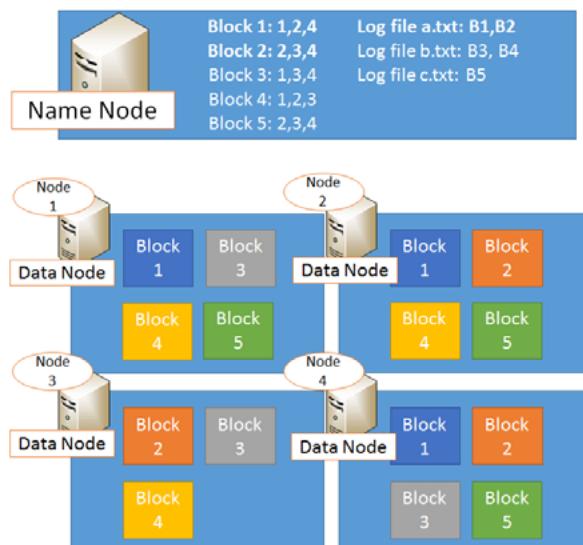
Em uma aplicação, os processos DataNode e JobTracker são multiplicados para cada máquina (instância) que participa da execução, enquanto o NameNode, JobTracker e SecondaryNameNode são criados uma única vez. Vamos conhecê-los:

**NameNode** – É o processo que tem a responsabilidade de gerenciar arquivos. Para um bom desempenho, suas informações ficam no nó mestre da aplicação localizada na memória RAM. Ele divide os arquivos de entrada em blocos, mapeia sua localização nos clusters, envia os blocos aos nós escravos, gera os metadados e controla sua localização, juntamente com suas cópias no sistema HDFS.

**DataNode** – O DataNode é efetivamente o responsável por armazenar os blocos de dados no HDFS. Ele é distribuído igualmente nos arquivos em diversas instâncias dentro da arquitetura Hadoop. Resumidamente, um DataNode é distribuído em diversas máquinas, armazena vários blocos de diferentes arquivos e informa o NameNode constantemente sobre quais blocos armazena e as modificações sofridas por eles.

A Figura 8 apresenta o NameNode com as informações do nó mestre e vários blocos gravados em DataNode de forma duplicada em cluster diferente.

Figura 8 – NameNode e DataNode



Fonte: Hadoop, 2019.

**SecondaryNameNode** – É o processo de apoio aos serviços do NameNode e atua na confiabilidade da arquitetura, quando há uma falha do NameNode, ele serve de recuperação. Trabalha com a função de *checkpointing* (ponto de checagem) em intervalo de tempo definido no NameNode.

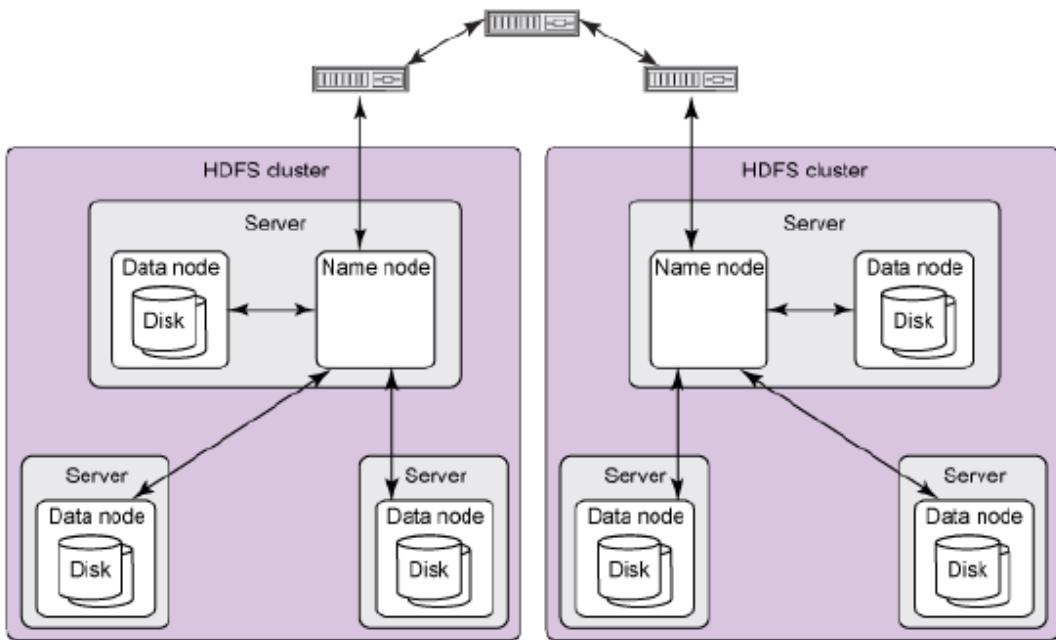
**JobTracker** – É o processo conhecido como o ponto de falha do serviço MapReduce do Hadoop, se ele parar, todos os demais serviços também são interrompidos. Sua função é gerenciar a execução das tarefas do MapReduce, pois cria as tarefas do MapReduce nos nós dos clusters, troca informação sobre localização dos blocos com o NameNode, localiza nós disponíveis e envia o trabalho para os nós do TaskTracker. Quando uma tarefa falha, o JobTracker pode reenviá-la para outro nó, marcá-la como não confiável ou sinalizar que o TaskTracker não é confiável. Ao final de seu processamento, todos os seus status de controle são atualizados.

**TaskTracker** – O TaskTracker é um nó de um cluster que operacionaliza as tarefas do JobTracker, executando o mapeamento, a redução e o embaralhamento. Com o objetivo de bom desempenho na sua operação, primeiramente, ele procura agendar a tarefa no *slot* do cluster em que ela se encontra junto com o DataNode com os dados; não sendo possível, ele procura por outro *slot* em uma máquina no mesmo cluster.

## 2.2 HDFS – Hadoop Distributed File System

O HDFS é o sistema de arquivo distribuído do Hadoop, armazena centenas de dados em formato de blocos pequenos de dados nos agrupamentos de servidores (cluster) distribuídos em milhares de nós da rede. Essa arquitetura de HDFS permite que a função MapReduce trabalhe escalonada com pequenos blocos de dados, característica fundamental do Big Data. A Figura 9 mostra a arquitetura do HDFS.

Figura 9 – Arquitetura do HDFS



Fonte: HANSON, 2013.

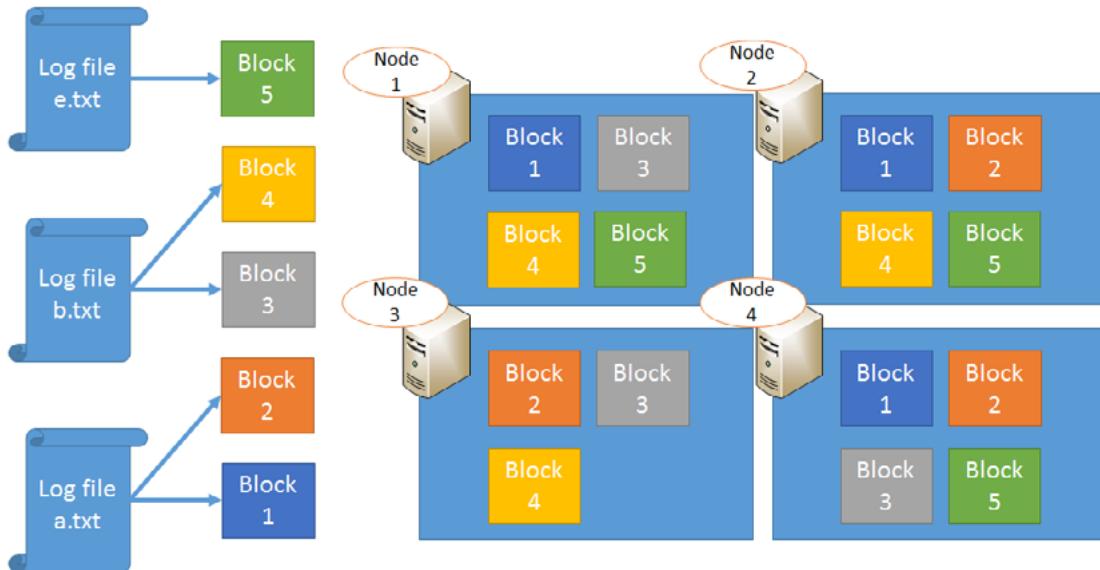
O HDFS trabalha com a ideia de localidade de dados, isso significa que servidores de baixo custo de processamento e armazenamento estão disponíveis em um grande cluster Hadoop. O desempenho é incrementado, pois o MapReduce Hadoop utiliza o mesmo servidor para armazenar e processar os blocos.

Um dos pontos fortes e elogiados do HDFS é a sua capacidade de tolerar e recuperar falhas. Lembramos que dados são quebrados e armazenados em blocos, esses são enviados e armazenados em clusters Hadoop espalhados pela rede, ou seja, um arquivo é particionado em blocos pequenos que são replicados em múltiplos servidores do cluster. Geralmente, um bloco é armazenado em três servidores e pelo menos dois blocos ficam em rack de servidor separado.

Mas qual o benefício? A arquitetura parece ser complexa. Algumas vantagens estão na redundância que oferece disponibilidade no cluster Hadoop, que pode quebrar o processamento em tarefas menores e executá-las em todos os servidores do cluster, obtendo, assim, o recurso

da escalabilidade. A Figura 10 apresenta como funciona a replicação de blocos de dados considerando quatro clusters no HDFS.

Figura 10 – Replicação de blocos no HDFS



Fonte: HANSON, 2013.

Observando a Figura 10, vamos analisar o segundo arquivo, o Log File b. O arquivo é separado em dois blocos de dados, sendo o número 3 e o número 4. Esses blocos são gravados em um cluster (*node 1*) e replicados em dois clusters adicionais que estão em rack físico separados. O bloco 3 é armazenado nos *nodes* 3 e 4. O bloco 4 é armazenado nos *nodes* 2 e 3.

Como o Big Data trabalha com grandes volumes de dados, o Hadoop também possibilita configurar o tamanho dos blocos em cada servidor, dessa forma, a performance pode ser ajustada. O desempenho do processamento em um cluster é suportado devido à capacidade de execução de grandes tarefas localmente, a transferência para outros cluster da rede é dispensada.

Na arquitetura, o servidor NameNode é responsável pela lógica de gerenciamento de dados, pelos arquivos do HDFS, localização de armazenamento dos blocos e outras. Como o NameNode é

responsável pelo gerenciamento de todo o conjunto de dados, é recomendado que o servidor escolhido para ele seja robusto, a fim de minimizar o risco de falha, e o procedimento de *backup* seja definido e acompanhado rigorosamente.

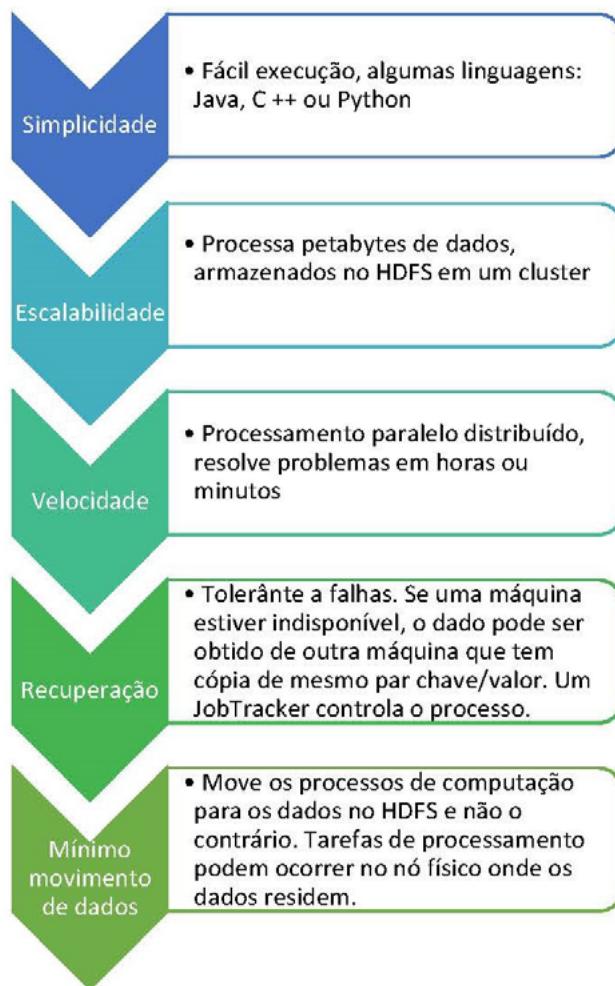
Na arquitetura Hadoop, o desenvolvedor não precisa se preocupar em saber onde está o dado armazenado, o Hadoop faz todo o gerenciamento e sabe onde gravar e recuperar. Quando a aplicação do desenvolvedor solicita a leitura de um dado, o Hadoop MapReduce acessa o NameNode, que localiza os servidores que contêm os dados e os envia ao aplicativo para execução no servidor local.

Para uma situação em que a aplicação grava um dado ou cria um arquivo, o HDFS faz a comunicação ao NameNode, que faz o armazenamento em servidor específico e a replicação do dado. Portanto, o NameNode gerencia os metadados do cluster que indicam a localização dos arquivos e os dados da aplicação processados pelo MapReduce nunca passam pelo NameNode

### ► 3. Modelo de programação MapReduce

A lógica do MapReduce é o processamento de enorme volume de dados de forma distribuída e paralela. Processa petabytes de dados, tanto estruturados como não estruturados, armazenados no HDFS. Faz parte da sua característica a simplicidade, escalabilidade, velocidade, recuperação e a movimentação mínima de dados. A Figura 11 apresentada abaixo resume essas características.

Figura 11 – Caracterização do MapReduce



Fonte: adaptado de Hortonworks (2019).

O processamento de dados realizado pelo modelo de programação MapReduce é feito com base na divisão e distribuição dos dados juntamente com um processamento paralelo distribuído e, finalmente, a redução dos dados e armazenamento.

Vejamos o processo: o MapReduce divide um grande conjunto de dados em vários blocos independentes (configuráveis em seu tamanho) e os distribui de forma organizada em pares de chave e valor para serem processados paralelamente. Esse processo paralelo é que dá velocidade e confiança a todo o trabalho.

A aplicação Map é que se responsabiliza pela divisão do conjunto de dados de entrada e cria tarefa de mapeamento de blocos para cada intervalo de entrada. Uma função, chamada Job Tracker, distribui nos nós de trabalho as tarefas de mapeamento de blocos dos intervalos, tendo como saída agrupamento de pares valores-chave para cada redução.

A aplicação Reduce é responsável por coletar os resultados, ou seja, o agrupamento de pares-valores espalhados, e fazer a combinação para responder ao problema principal. Isso é possível, pois a cada execução de Reduce, uma parte relevante dos dados mapeados pela Map é executada e seu resultado é gravado novamente no HDFS. Concluindo, o Reduce resolve um problema coletando os dados do mapeamento, por meio da chave, e combinando seus valores.

## ASSIMILE

Como o Hadoop possui uma arquitetura flexível, outros aplicativos podem fazer parte da arquitetura, a isso chamamos de ecossistema de aplicativos do Hadoop. Exemplo: Hive (armazenamento de dados com linguagem HiveQL), HBase (software de banco de dados, transforma tabelas em formulário de entrada e de saída do MapReduce) e Pig (plataforma com linguagem para analisar dados).

## ► 4. Componentes do Hadoop

O Hadoop tem como seus principais componentes o sistema de arquivo distribuído HDFS e o modelo de programação MapReduce. Devido ao seu desempenho ser aberto para uso, ele facilmente foi aceito pela comunidade da computação, que desenvolve novos projetos com

propostas adaptadas a algumas necessidades de mercado. Essas propostas, chamadas de projetos ou subprojetos, são incorporadas ao *framework* da arquitetura do Hadoop, fazendo com que ele se adapte e se torne mais completo. Apresentaremos alguns dos projetos de forma resumida, não trataremos do HDFS nem do MapReduce, pois já foram discutidos. Lembrando, ainda, que há outros projetos relacionados a arquitetura Hadoop que não serão citados.

**Hadoop Common Components – É uma ampla lista de comandos** agrupados em bibliotecas. Tem como objetivo auxiliar os projetos e as interfaces de outros sistemas de arquivos. São exemplos: ferramentas de manipulação de arquivos, permissão de arquivos, classificação de dados, etc.

**Linguagem de desenvolvimento de aplicações** – Muitas linguagens de 3<sup>a</sup> e 4<sup>a</sup> geração são utilizadas no modelo de programação do Hadoop, exemplo: Pig e Jaql.

- PIG – Tem o objetivo de analisar grandes conjuntos de dados, por meio de execução paralela. É uma linguagem de programação de alto nível desenvolvida pela Yahoo. Possibilita criar programa de bom desempenho nas execuções de mapeamento de redução.
- Jaql – É uma linguagem que teve seu início na Google, mas que a IBM adotou em seus pacotes de Hadoop. Permite processamento e consulta de dados em arquitetura Hadoop. É flexível no acesso à base de dados, suporta JSON, XML, CSV e outros. Possui bom desempenho e é recomendada para vários usos quando comparada ao PIG e HIVE.

**HYVE** – São softwares abertos para o uso. Eles manipulam volume grande de dados e utilizam a linguagem HiveQL para consulta e análise desses dados. Sua interface HiveQL é semelhante ao SQL, portanto, faz consulta e escrita em tabelas do HDFS de forma distribuída.

**Hadoop Stream** – A arquitetura permite, além da linguagem Java, que a gravação de mapeamento e redução dos blocos de dados sejam realizadas por outras linguagens. O acesso a esses dados também podem ser feitos por meio de APIs, conhecidas como Hadoop Streaming. O padrão de entrada e saída de uma API Streaming é que faz a interface com a aplicação. As interfaces de *stream* são simples e pequenas e podem ser desenvolvidas com Python ou Ruby. Resumidamente, o *stream* permite que os diversos dados das atividades on-line na internet sejam organizados de forma comum e disponibilizados de várias formas para pesquisa e análise.

**Avro** – Este projeto provê serviço em JSON para estruturar e serializar dados. É executado remotamente em formato compactado e binário. Os dados gravados pelo serviço remoto (RPC) do Avro segue um esquema de organização dos dados. Esse esquema também é armazenado com eles. O esquema é utilizado para que aplicações diversas possam acessar e identificar os dados posteriormente.

**Hbase** – É um software de banco de dados orientado à coluna. Faz manipulação em tempo real em tabelas distribuídas, lê e grava aleatoriamente dados em enormes tabelas (bilhões de linhas e milhões de colunas). Ele é executado no topo do HDFS e seu modelo segue o armazenamento distribuído da Google, o Bigtable.

**Flume** – É um canal de direcionamento de fluxo de dados. O sistema opera por meio de processamento distribuído na ação de coletar, juntar e mover dados em diversos sistemas de arquivos, com o objetivo de armazená-los em sistema centralizado.

**Lucene** – É um projeto que realiza pesquisa de texto. É formado por um conjunto de biblioteca de engenharia de busca de texto de alto desempenho, recomendado para aplicativos que desejam realizar pesquisa de texto.

**ZooKeeper** – Provê a centralização de infraestrutura e serviços para sincronização e coordenação de tarefas das diversas aplicações no cluster, dentre elas: a configuração de nós, hierarquia, nome do nó, sincronização de processos e outras.

## TEORIA EM PRÁTICA

O Hadoop é importante atualmente, pois pode auxiliar as organizações nas suas decisões baseadas em dados em tempo real, pois possibilita execução de atividades com vários formatos de dados, sejam eles relacionamento e sentimento de mídia social, fluxo de cliques, *streaming* de vídeo e áudio e outros, podendo ser também um dado semi e não estruturado. Também impulsiona o futuro da ciência de dados, um campo interdisciplinar que combina aprendizado de máquina, estatística, análise avançada e programação. Permite transformar os dados do data warehouse local para um armazenamento distribuído baseado em Hadoop, consolidar os dados em toda a organização para aumentar a acessibilidade, diminuir os custos e acelerar as decisões com mais precisão (IBM, 2019).

Faça uma reflexão sobre a importância da arquitetura Hadoop, se necessário, realize nova leitura do material de aula.

## VERIFICAÇÃO DE LEITURA

1. A plataforma Hadoop é mais indicada para análise de dado semiestruturado e não estruturado ou quando se deseja atender a uma necessidade ou resolver um

problema não específico. Considerando a arquitetura Hadoop, é INCORRETO afirmar:

- a. O Hadoop também pode ser utilizado para tratar dados estruturados.
- b. Os dados armazenados nos sistemas tradicionais passam por um rigor de tratamento até chegar à base de dados que será analisada.
- c. O Hadoop foi desenvolvido em linguagem Java e é um sistema de arquivos distribuído em cluster para computação em grande escala.
- d. O Hadoop foi projetado para receber tudo da internet, com a intenção de distribuir o que já é conhecido sem necessidade de análise.
- e. A arquitetura do Hadoop trabalha para produzir resultados com base na alta escalabilidade e processamento *batch* distribuído.

2. De acordo com a arquitetura Hadoop, qual afirmativa abaixo é correta?

- I. O MapReduce quebra em tarefas um trabalho, realiza processamento paralelo massivo pela rede e reduz a tarefa em novas tarefas para manipular dados e armazenar em servidores.
- II. A arquitetura do Hadoop digitaliza dados com o objetivo de produzir resultados com base na alta escalabilidade e sistema de processamento *batch* distribuído.
- III. Não é um projeto que visa velocidade de tempo de resposta ou armazenamento em tempo real, velocidade de armazenamento. Foi projetado para busca e descoberta, com a intenção de fazer algo desconhecido.

IV. A arquitetura Hadoop é formada somente pelo HDFS – Hadoop Distributed File System.

- a. I e III.
- b. II e III.
- c. I, II e III.
- d. Somente a II.
- e. I e II.

3. O Job Tracker não é eficaz na gerência da execução das tarefas do MapReduce e não é capaz de criar as tarefas nos nós do cluster.

POR TANTO

O JobTracker é conhecido como o ponto de falha do serviço MapReduce do Hadoop. Se ele parar, todos os demais serviços também são interrompidos.

Com base nas informações dadas e na relação proposta entre elas, é CORRETO afirmar que:

- a. As duas afirmações são verdadeiras e a segunda justifica a primeira.
- b. As duas afirmações são verdadeiras e a segunda não justifica a primeira.
- c. A primeira afirmação é verdadeira e a segunda é falsa.
- d. A primeira afirmação é falsa e a segunda é verdadeira.
- e. As duas afirmações são falsas.

## ► Referências bibliográficas

HANSON, J. J.; **An introduction to the Hadoop Distributed File System.** Explore HDFS framework and subsystems. IBM. 2013. Disponível em: <<https://www.ibm.com/developerworks/library/wa-introhdfs/index.html>>. Acesso em: 30 ago. 2019.

HADOOP. Disponível em: <https://hadoop.apache.org/docs/r3.1.2/index.html>. Acesso em: 24 mai. 2019.

HORTONWORKS. Disponível em: [https://br.hortonworks.com/apache/mapreduce/#section\\_1](https://br.hortonworks.com/apache/mapreduce/#section_1). Acesso em: 27 mai. 2019.

IBM. **O que é o Apache Hadoop.** Disponível em: <https://www.ibm.com/br-pt/analytics/hadoop>. Acesso em: 24 mai. 2019.

STRIEN, K. V. **Hadoop:** HDFS en MapReduce. Disponível em: <https://blogs.infosupport.com/hadoop-hdfs-en-mapreduce/>. Acesso em: 29 mai. 2019.

WHITE, T. **Hadoop – The Definitive Guide.** 3. ed. California, USA: O'Reilly Media, Inc., 2012.

## ► Gabarito

### Questão 1 – Resposta D

**Resolução:** a afirmação D está INCORRETA, pois, ao contrário do que é afirmado, o Hadoop foi projetado para busca e descoberta, com a intenção de fazer algo desconhecido ou impossível se tornar próximo de uma possibilidade de análise.

### Questão 2 – Resposta C

**Resolução:** as afirmativas I, II e III estão corretas. A afirmação IV está incorreta, pois a composição básica do Hadoop é o HDFS – Hadoop Distributed File System e o modelo de programação MapReduce.

### Questão 3 – Resposta E

**Resolução:** a primeira afirmação é falsa, pois a responsabilidade do Job Tracker é gerenciar a execução das tarefas do MapReduce, pois cria as tarefas do MapReduce nos nós do cluster, troca informação sobre localização dos blocos com o NameNode, localiza nós disponíveis e envia trabalho para os nós do TaskTracker. A segunda afirmação é verdadeira, pois trata-se de um ponto de segurança do JobTracker.



# **Aplicação da ciência de dados no gerenciamento empresarial**

Autor: Aimar Martins Lopes

## **► Objetivos**

- Compreender como a ciência de dados contribui para as organizações.
- Associar a estrutura organizacional e a inteligência dos algoritmos ao Big Data.
- Conhecer e avaliar o modelo de negócio baseado em dado.
- Avaliar a importância do controle e o acesso aos dados.

## 1. A importância do Data Science

Poucas pessoas sabem e percebem que as organizações empresariais e seus negócios, bem como a sociedade como um todo, estão envolvidos com dados e informações. Os indivíduos, os dispositivos digitais e as organizações geram um fluxo de informação gigantesco em todo o planeta. O nosso assunto se refere a isso, os dados gerados pela sociedade e suas relações com os negócios das organizações.

A maioria dos dados que estão sendo armazenados são provenientes dos últimos anos, a tendência é de crescimento vertiginoso para os próximos anos. A conclusão desse cenário é que temos muitos dados à disposição e a análise, além de necessária, torna-se um grande negócio. Contudo, no meio desse caminho, há a questão da privacidade do indivíduo, o controle e o acesso aos dados. Esse é o principal motivo e importância do nosso estudo. Portanto, qual é a relação da ciência dos dados ou Data Science com os negócios e a sociedade?

Você está pronto? Vamos em frente.

Atualmente, o assunto Data Science e *Big Data* está em boa parte das agendas dos executivos, tão relevante que o mercado não hesita em afirmar que o dado é tão importante que compará-lo à importância do petróleo para a sociedade em décadas passadas não é exagero, por isso, diz-se que o dado é o petróleo da era digital. Portanto, o dado passa a ser um ativo importante, tanto para a organização em si como para aqueles que veem uma oportunidade de ganhar dinheiro por meio de equipamentos, software, modelo matemático e estatístico, algoritmo, entre outras coisas. Porém, sem um modelo sistêmico, um método e recursos adequados, o uso dos dados não tem efeito. São necessários métodos de análise efetivos que criam *insights*, mostrem significados e auxiliem na tomada de decisão. É aí que entra o Data Science, que vem para ajudar a resolver problemas organizacionais no mundo dos dados.

Figura 12 – Integração dos dados



Fonte: DrAfter123 / iStock

## ASSIMILE

Você encontrará várias definições de **ciência de dados**. Um bom conceito está fundamentado em conhecimento interdisciplinar, utilizado para coleta e análise e visualização de dados por meio de diversas técnicas, com o objetivo de criar *insights* para tomada de decisão

## ► 2. O mundo dos dados

Para a comprovação de algo, é necessário a apresentação de dados ou número que comprove o fato. Contudo, esses mesmos dados podem ser enigmáticos e, apesar de provar algo, pode não ter o significado

esperado. Podemos dizer que o dado possui várias faces. Por exemplo, se uma rede de supermercado teve uma explosão de venda de um determinado produto, não significa que os clientes gostem desse produto, a venda pode ter sido motivada por uma promoção ou uma moda assimilada pelos clientes. Conseguirá vantagens, ou seja, bons *insights*, aqueles que mais corretamente conseguirem analisar e interpretar os dados dentro de um contexto ou uma relação.

Apesar de os dados representarem situações de clientes, produtos e outras coisas, eles estão totalmente abertos para a interpretação. Como as tecnologias associadas a *Big Data* crescem rapidamente e se infiltram nos negócios das organizações, os executivos necessitam de recursos para entender de forma aguda o significado da análise dos dados e os números resultantes. No mundo dos negócios, atualmente, o constructo valor, ou criação de valor, que de forma simples significa como algo impacta o cliente fazendo ele retornar, é o dever de casa que as organizações precisam realizar e responder, por meio da análise de dados, como seus produtos e serviços podem criar valor na sua cadeia de atuação, e claro, para o cliente.

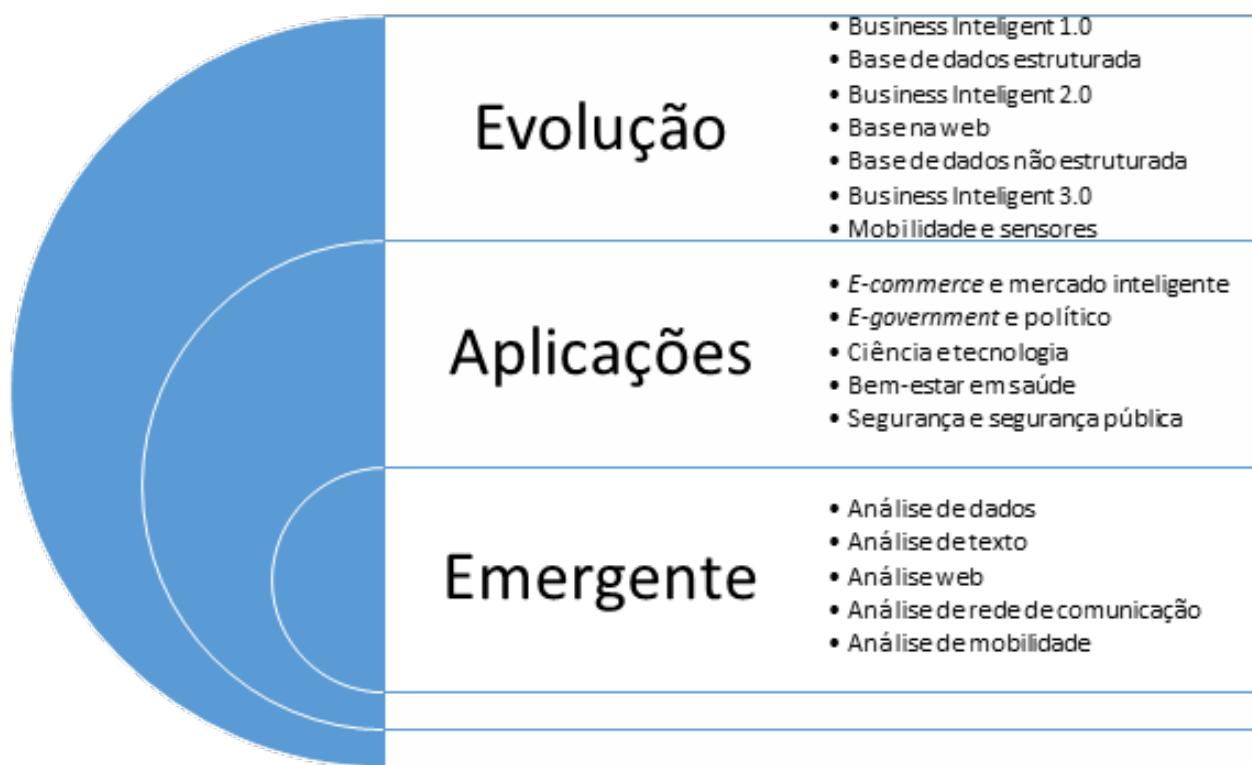
O tipo e a quantidade de dados são importantes, mas estabelecer uma cultura corporativa sistematizada e profunda de análise de dados para a tomada de decisão sobre evidências é fundamental. O resultado será uma fonte indiscutível de dados valiosos, com *feedback* real para os tomadores de decisão, mudanças nas regras de negócios, entre outras decisões. A adoção sistematizada do uso eficaz dos dados irá encorajar outras pessoas na organização a trabalhar da mesma forma. (O'CONNELL, 2014).

O conceito de ambiente organizacional que envolve informações internas e externas de uma determinada indústria mostra o crescente aumento de informações exógenas, contrapondo as informações endógenas estruturadas. As informações exógenas não estruturadas provêm de redes sociais, e-mail, mídias de colaboração em massa,

dispositivos de IoT, tráfego de internet, fotografias, gráficos, entre outros. Essa quantidade de dados forma o Big Data, caracterizado por volume, variedade, velocidade, veracidade e valor. (DAVENPORT, 2014).

A ciência de dados no campo da análise fornece uma realidade de alto impacto na sociedade, com muitos desafios, ameaças e oportunidades. A Figura 13 abaixo apresenta a evolução do Business Inteligente (BI) e do Analytics (análise de dados), inclui aplicativos e oportunidades de análise.

Figura 13 – BI e análise de Dados Evolution, Applications, and Emerging Research



Fonte: adaptado de Chen et al. (2012).

O termo “inteligência” tem sido utilizado desde 1950. Business Intelligence tornou-se popular por volta de 1990, e nos anos 2000, Business Analytics foi introduzido para representar o fator-chave da análise do BI. Mais recentemente, Big Data e Big Data Analytics tornou-

se comum para se referir a *data sets* (conjunto de dados) e *analytical techniques* (técnicas de análise) em grandes e complexos conjunto de dados que requerem avançadas tecnologias de armazenamento, gerenciamento, análise e visualização (CHEN et al., 2012).

A Figura 13 apresentada anteriormente representa a evolução do BI e a análise de dados está dividida em:

- Evolução, caracterizada pelo armazenamento de dados em bancos de dados estruturados evoluindo para bases não estruturadas com influência da web e atualmente com tecnologias de mobilidade e sensores;
- Aplicações abrangem praticamente todas as áreas, comércio eletrônico, governo, política, pesquisa, saúde e segurança;
- Pesquisa emergente, composta por análise de Big Data, web, texto, rede e mobilidade.

### ► 3. A relação Big Data e negócios

Alguns estudos científicos e organizacionais têm se preocupado em conhecer como o Data Science (ciência de dados) e o Big Data estão relacionados aos negócios e à sociedade.

A análise dos dados para obter informações que subsidiem a tomada de decisão organizacional possui características um pouco diferentes de captação, processamento e apresentação. Palavras e expressões são utilizadas para expressar relações e dimensões de conceitos, nas áreas de negócios e tecnologia às quais o tema Big Data está relacionado não é diferente. Portanto, para uma unificação do termo, trataremos o assunto com o uso da expressão Big Data sempre que necessário. Para compreender melhor as palavras e expressões envolvidas com Big Data

e Data Science, apresentamos a Figura 14: as palavras contidas nela dão uma dimensão melhor da abrangência do termo, pois, quanto mais próximo da palavra e maior, mais significado tem.

Figura 14 – Palavras relacionadas com Big Data e Data Science



Fonte: Mindscanner / iStock

# PARA SABER MAIS



Caso goste de analisar ou deseje criar um modelo de negócios, pois o mundo digital demanda inovações na forma de criar valor para o cliente e fazer negócios, consulte a obra *BMG – Business Model Generation: inovação em modelos de negócios*, de Alexander Osterwalder e Yves Pigneur. Você irá aprender de forma prática e inovadora como criar ou aprimorar um modelo de negócio. Esse método é utilizado por grandes corporações e startups inovadoras em todo o mundo.

O estudo de Silveira et al. (2015), como resultado, apresenta um conjunto de palavras que qualifica várias áreas e conceitos de Big Data. As relações das palavras, expressões, envolvem tecnologia de sistemas de informação e áreas de negócios. Portanto, reúnem dois grupos: Big Data e negócios. Essas relações direcionam a incidência para áreas emergentes da atualidade, tais como: redes e mídias sociais, armazenamento e análise de grande volume de dados (*data warehouse*), mineração de dados (*data mining*), qualidade dos dados, e o desafio para o indivíduo, diante dessa oportunidade, é entender e contextualizar tudo isso.

Conclui-se que há um forte elo entre o gerenciamento da informação e a tomada de decisão. As conclusões mostram que os autores defensores do crescimento do interesse da sociedade, das pessoas e dos negócios na relação Big Data e negócios estão corretos (SILVEIRA et al., 2015).

É de muito tempo atrás a preocupação das organizações com a questão do uso dos dados, o principal objetivo é utilizar os dados para melhorar seu desempenho. Contudo, somente agora temos à disposição tecnologia capaz de produzir, transferir, processar, armazenar e disponibilizar informação de forma eficiente e eficaz. No mundo dos negócios, as áreas que mais utilizam dados estão relacionadas no Quadro 1.

Quadro 1 – Áreas de negócios com potencial de uso de dados

Uso intenso de dados	Alto potencial de uso de dados
<ul style="list-style-type: none"><li>• Empresas ".com" (on-line)</li><li>• Atacadistas</li><li>• Seguradoras</li><li>• Turismo</li><li>• Bens de consumo</li><li>• Logística</li><li>• Segurados</li><li>• Bancos e cartões de crédito</li></ul>	<ul style="list-style-type: none"><li>• Bancos</li><li>• Mídia e entretenimento</li><li>• Energia</li><li>• Telecomunicação</li><li>• Indústria de manufatura</li><li>• Varejo</li><li>• Saúde</li><li>• Pequenos negócios</li></ul>

Fonte: elaborado pelo autor.

A tendência nesse assunto está direcionada para disponibilidade cada vez maior de recursos para realizar o processo dos dados com menor custo e mais ferramentas gratuitas, portanto, por diversos motivos, o acesso ao dado e sua análise para transformá-lo em informação estará disponível para diversos meios de produção, organização, mesmo as menores e, inclusive, para profissionais liberais.

Outra tendência diz respeito a características de volume do Big Data, ela está relacionada com grande quantidade de dados que na maioria das vezes serão genéricos, mas que podem conter um caráter de especificidade. Por exemplo, o time de beisebol americano Giants utiliza durante uma partida os dados trocados pelos torcedores na rede social do time. A intenção é de melhorar a experiência e o engajamento do torcedor com o time e com o evento. Essa é uma situação bem específica.

Outros exemplos na área de marketing e atendimento ao cliente podem criar *insight* com o objetivo de personalizar o produto ou serviço. Para Huang e Van Mieghem (2014), quanto mais capacidade analítica, mais informações úteis as organizações têm de seus clientes, geram mais boca a boca e reduzem o custo.

Com a tendência do aqui e agora, a tomada de decisão deve ser em tempo real ou o mais próximo disso possível. Para atender a esse requisito, a organização necessidade de coleta e análise de dados online. A análise em tempo real de uma fila de espera ou atendimento é um exemplo desse modelo, outro é o fluxo de clique de um indivíduo pelo site: associado com sugestão baseada no comportamento desse indivíduo, pode-se ainda associar premiação num programa de bônus quando um cliente aceita uma sugestão ou compra algo proveniente de uma campanha, entre outras coisas.

O sucesso da organização, provavelmente, estará na personalização, na inovação e na capacidade de tomar decisão em tempo real, bem como

adaptar ou implantar estruturas de serviços e produtos que criem valor para o cliente.

## PARA SABER MAIS

Para se aprofundar na questão dos dados é importante pesquisar “Big Data e suas tecnologias”, pois Big Data não é uma tecnologia única e específica, trata-se de um conjunto de práticas de gestão de dados apoiado em diversas tecnologias de *analytics*. O *mindset* a ser formado deve considerar a transição de noções simples de dados e análise para solução de problema e inovação específica de um negócio.

## ► 4. Debates sobre a análise de dados

Se pesquisarmos sobre os desafios das organizações para a análise de dados e o Big Data, na tentativa de criar valor por meio da prática, geração de modelos de dados e controle dos supostos *insights*, encontraremos algumas abordagens interessantes, seguem algumas delas.

**Big Data na prática** – Este campo se refere a como os indivíduos tratam os dados na prática, ou seja, no dia a dia dos processos organizacionais. Abordagens indutivas e dedutivas da análise dos dados que dão um bom conhecimento baseado na origem de fontes diferentes de dados, tais como: sistemas de ERP, processos internos, fontes externas, dados de terceiros, dados abertos, sensores e rede sociais.

A consequência dessa abordagem é que teremos dados produzidos e gerados aleatoriamente, não são dados produzidos para objetivos específicos, podemos chamar de dados eventuais. Dependendo da granulidade e variedade desse modelo, pode-se ter dificuldade para *insight*. A Figura 15 mostra a variedade das tecnologias que podem ser envolvidas dentro do mundo dos dados.

Figura 15 – Ciência de dados e tecnologias participantes



Fonte: nongpimmy / iStock (c).

**Estrutura centralizada, descentralizada ou híbrida para os dados** - A organização deve ter a capacidade de criar uma estrutura para lidar com a análise de dados, necessita contar com formas de desenvolver e usar os recursos humanos e técnicos relacionados a Big Data. O mercado oferece inúmeras formas de coletar, acessar, rastrear, gerir, processar, analisar e dispor os dados para a tomada de decisão.

A dificuldade não está somente em adquirir ou desenvolver os recursos, mas também em como dispor esses recursos e como estruturar a equipe. A centralização é uma alternativa, a criação de centros de análise de dados com capacidade analítica tem sido considerada por algumas organizações, pois reúne um centro de inteligência e ao mesmo tempo reduz a escassez de competências analíticas. Nesse centro têm-se competência de análise de dados e experiência de negócio com a função de atuar como prestador de serviço para as unidades de negócios da organização. Uma boa vantagem está na autonomia de decisão sobre os processos e a governança relacionada a análise dos dados, bem como a segurança.

Uma crítica à estrutura centralizada para a organização dos dados é fundamentada na necessidade de existir uma diretoria para fazer a governança da área, cada cargo criado aumenta os custos de estrutura, comunicação, etc., e a relação da estrutura de dados tem dificuldade na colaboração e ação sinérgica entre as unidades de negócios.

Uma estrutura descentralizada é constituída por um ambiente mais propenso ao compartilhamento de informação e criação de *insights* de valor aos negócios, além de potencializar a comunicação e o envolvimento de diferentes *stakeholders*. Alguns casos de desempenho superior em projetos de dados mostram que os fatores de sucesso de um projeto de análise de dados estão associados com equipes multidisciplinares, embora o amadurecimento tenha exigido cada vez mais a execução da análise pelo tomador de decisão. Por outro lado, o analista também precisa conhecer e se engajar mais nos negócios da organização, ou seja, precisa estar próximo aos negócios.

No entanto, as organizações são formadas por complexos processos que se modificam de acordo com as variáveis ambientais, portanto, um arranjo de estrutura híbrida, entre o centralizado e o descentralizado, pode ser viável. Pode-se centralizar a organização, governança e

segurança e descentralizar profissionais especializados para atuarem próximos às unidades de negócios.

**Inteligência dos algoritmos** – O analista de dados ou cientista de dados responsável pela análise e interpretação dos dados deve tomar muito cuidado para não ser influenciado por viés de comportamento, perfil da equipe, crença, ceticismo, pressões de tempo e custo, comparação de dados de entrada e saída, visualização e, por fim, decisão baseada em evidência. Então, o que fazer para minimizar, evitar ou controlar essas influências?

Os cientistas de dados e profissionais da área pensam que a solução está em investir e explorar os algoritmos inteligentes capazes de análise de grandes volumes de dados e em tempo real. As organizações estão cada vez mais investindo e criando processos baseados nesses algoritmos, isso justifica a popularidade dos algoritmos inteligentes.

Para Günther et al. (2017), a tendência é para o uso de algoritmos inteligentes, pois esses criam uma trilha de conceitos analíticos inovadores e sistemáticos, evitando, assim, as influências provenientes de preconceitos já estabelecidos. A lógica algorítmica processa os dados por meio de parâmetros fixos e pré-programados, mas o uso de modelos associados aos algoritmos leva a padrões e *insights*.

Diversas organizações investem em modelos de negócios, produtos, serviços, equipamento, entre outras inúmeras coisas que são baseadas em algoritmos inteligentes. Exemplo: redução do tempo logístico, fraudes em seguro e contratos, carro autônomo, modelo de tráfego para cidades inteligentes, recomendação de pessoas ou negócios. Na área de humanas, os algoritmos de análise de comportamento estão cada vez melhores, além de realizarem análises descritivas, fazem também predição. Por outro lado, os algoritmos, quando considerados para tomada de decisão estratégica, são mais complexos e geralmente

necessitam da intervenção humana, apesar dos diversos algoritmos que tomam decisões sem a interferência humana.

### **PARA SABER MAIS**



Dois conceitos estatísticos devem ser considerados numa análise de dados:

- **Descrição** é a avaliação e, consequentemente, um relato ou uma narração detalhada sobre os dados (série temporal) por meio de característica, enumeração, média, etc.;
- **Predição** é um conjunto de métodos que, na análise dos dados, tem como ideia principal a existência de padrões específicos intrínsecos difíceis de encontrar e interpretar, mas que irão se repetir em um dado momento. A predição também procura informar as possíveis variações no decorrer de outras análises.

## **5. Modelo de negócio baseado em dados**

Um modelo de negócios é a forma como uma organização está arranjada ou representada para criar valor na sua indústria de competição, geralmente os modelos de negócios são resultados da definição do posicionamento estratégico, considerando a visão, missão, objetivos e outros elementos.

Dada a importância dos dados, eles vêm sendo considerados um elemento influenciador mandatário do modelo de negócios. As organizações já estabelecidas no mercado por um longo tempo têm

dificuldade em flexibilizar e inovar seus modelos de negócio, a barreira cultural e a equipe madura dificultam a mudança.

No entanto, as organizações *startups* conseguem definir modelos de negócios inovadores baseados em dados e encontram poucas barreiras de entrada nas indústrias. Exemplos mais conhecidos são: Netflix, Uber e Airbnb.

Para superar a barreira da inovação, as organizações mais tradicionais têm procurado disponibilizar para seus colaboradores novas formas de usar e analisar os dados, ações de melhoria de processos, compartilhamento da base de dados. Essas ações visam manter o funcionamento normal dos negócios e aproveitar os dados de forma mais eficaz.

Os modelos de negócios baseados em dados irão substituir os modelos tradicionais menos eficientes, pois esses modelos proporcionam a criação de novos valores, descobrimento de novos mercados, formas de engajamento, segmentação de nicho e aumentam a interação com o cliente.

Um modelo de negócios baseado em dados quebrou o paradigma tradicional do mercado de locação de filmes para residência. Esse modelo foi implantado pela Netflix, que não só saiu do aluguel de disco e filme VHS para um modelo de *streaming* utilizando a internet como também produz conteúdo baseado em dados, faz recomendação dinâmica de acordo com o perfil e comportamento do cliente. Nesse modelo, a Netflix triplicou sua base de assinantes até 2016. Outros exemplos são a IBM, com a plataforma Watson; a Nike, que além de fabricante de calçados, possui uma plataforma de serviços de condicionamento de dados; e a Starbucks, que desenvolve novos produtos com base nas sugestões dos clientes.

Resumidamente, as organizações modificam e criam modelos de negócios para monetizar dados ou *insights*. Quando entregam uma novidade ou aprimoram a experiência do cliente, elas entregam proposições de valor antes desconhecidas por ele. Como vimos nesses exemplos, concluímos que o Big Data possibilita alteração radical das estratégias dos negócios.

Na medida que a organização aumenta sua maturidade em Big Data, seu desempenho melhora, não se tratando de uma implementação única e rápida com resultados imediatos, e sim de um crescimento paulatino. A organização deve primeiro idealizar e implantar a infraestrutura e obter excelência operacional, só assim conseguirá elaborar novos modelos de negócios. Contudo, o sucesso da mudança para uma empresa estabelecida, mesmo com grandes recursos, também depende do tipo de indústria e do tamanho da organização.

## ASSIMILE

Um **modelo de negócio** é uma forma clara na qual uma organização atende a um segmento de cliente, proporcionando um conjunto de valores por ele reconhecido. Nessa relação, devem estar claros o produto, o valor e os recursos necessários, como também a relação de entrada e saída monetária do negócio.

## ► 6. Controle e risco dos dados para a organização

### 6.1 Controle de acesso aos dados

A segurança dos dados, bem como a democratização do acesso, seja ele interno, para os membros das organizações, ou externo, para a rede

de parceiros, como também a obtenção de dados oriundos de base de dados de livre acesso são questões relevantes na política de governança de dados a serem considerados pelas organizações e governos.

Para obter benefícios, as organizações devem manter um grau bom de confiança para a troca efetiva de dados com sua rede de parceiros, desenvolver boas práticas de triagem, acesso e divulgação. As regras que regem essa questão dependem da extensão do acesso aos dados e do controle de abertura ao tipo de dado disponível. Resistências a troca de dados entre parceiros é justificada pela segurança e privacidade, mas o fator principal reside na análise. Essa prática é percebida como ato potencial de concorrência, sua prática enseja ameaça à posição estratégica da organização.

Esse contexto leva à obrigatoriedade do controle do acesso aos dados pelas organizações. O Quadro 2 apresenta uma lista de ações para reduzir o risco do compartilhamento de dados.

## Quadro 2 – Formas de controle de acesso aos dados

### CONTROLE DE ACESSO AOS DADOS

- Contrato formal entre as partes envolvidas
- Explicitar o direito e a propriedade do dado
- Possibilidade ou não do uso para outros produtos
- Quem controla os dados e o acesso
- Regras de hierarquia de acesso aos dados
- Registro de acesso aos dados e auditoria
- Regras de coleta de dados
- Definir os canais fornecedores de dados
- Definir regras de *compliance* para os parceiros
- Definir quais dados poderão ter livre acesso para a sociedade
- Quais são as formas permitidas de replicação dos dados
- Qual a tratativa e responsabilidade pelos dados dos participantes

Fonte: elaborado pelo autor.

A organização deve lembrar que coletar, armazenar e processar dados tem consequências éticas e responsabilidade regulamentada por lei.

## 6.2 Os riscos associados ao valor do Big Data

A segurança dos dados não implica somente a falha no controle de quem acessa os dados, a análise combina uma série de informações de várias fontes pessoais de dados e pode revelar informações e comportamento peculiares de um indivíduo; o risco social aumenta muito mais para as situações em que as informações são vazadas. Se considerarmos que as organizações e a sociedade, de uma forma geral, tratam o Big Data e a análise de dados como uma fonte de valor preciosa, elas devem perceber que esse mesmo valor frequentemente aumenta os riscos sociais.

O comportamento organizacional tende à dedicação e à orientação estratégica para a coleta e análise dos dados, com o intuito de gerar produtos e serviços cada vez mais personalizados à cadeia de consumidor, portanto, os riscos inerentes aumentam as preocupações no que diz respeito a roubo de dados, discriminação ilegal, privacidade, análise equivocada de perfil entre outros.

Por um lado, o governo e a organização alegam que o controle de dados está associado à vigilância dos indivíduos e à concepção de uma melhor política de segurança pública; por outro, o indivíduo e a população sentem que seus limites de liberdade, autonomia e privacidade são reduzidos e invadidos.

Consideremos ainda as indústrias específicas que lidam com dados confidenciais, como a saúde, incluindo seus serviços médicos, laboratoriais, diagnóstico, tratamento, etc. Nessa área, o Big Data cresce consideravelmente, promovendo vários benefícios para todos os *stakeholders*, no entanto, não se vê avanço na extensão de medidas para cumprir as regulamentações ou o uso ilegal ou antiético do dado para evitar danos à reputação.

Além do controle de acesso aos dados, a governança empresarial pode tomar uma série de medidas para garantir o acesso, compartilhamento e uso dos dados de forma legal e ética. Seguem algumas medidas no Quadro 3.

### **Quadro 3 – Acesso e compartilhamento de dados**

ACESSO E COMPARTILHAMENTO DE DADOS
<ul style="list-style-type: none"><li>• Determinar objetivos de uso dos dados</li><li>• Autorização de uso com propósito definido</li><li>• Práticas de governança e transparência</li><li>• Política de retenção, segurança e descarte de dados</li><li>• Consentimento do indivíduo para compartilhar</li><li>• Controle de login de acesso</li><li>• Regras de transparência</li><li>• Consentimento claro do indivíduo</li><li>• Mecanismo de controle de acesso</li><li>• Explicitar o direito e a propriedade do dado</li></ul>

Fonte: elaborado pelo autor.

Contudo, encontramos situações em que as metas organizacionais estão acima da privacidade dos usuários, a estratégia é considerar minimamente ou ignorar a privacidade do indivíduo. Nesse caso, a organização esconde sua política de uso de dados, não permite que o indivíduo opine ou autorize a coleta e o compartilhamento de seus dados. Por exemplo, a organização em seus termos explica que o dado é coletado e compartilhado para determinados fins, no entanto não solicita autorização explícita do indivíduo.

Em países onde as leis de segurança de dados são incipientes, muitas organizações oportunistas adotam a estratégia de coleta e exploração dos dados e vão até o limite em que enfrentam uma ação legal. São inescrupulosas e sem consideração pelo indivíduo. Esse tipo de organização pensa que a privacidade é um problema e não

uma oportunidade para construir algo melhor, coletam e usam dados de forma não ética até que os indivíduos afetados imponham uma resistência maciça à sua prática, seja por meio de denúncias ou ações judiciais.

Todavia, há organizações que buscam o equilíbrio, coletam informações para seus negócios, mas não compartilham e nem vendem, como também permitem que os indivíduos proibam serviços baseados em dados pessoais.

Portanto, diante dos conceitos sobre a aplicação da Ciência de Dados no gerenciamento organizacional, podemos constatar que coleta e análise de dados possui um potencial gigantesco de negócios, há várias formas de uso dos dados, porém a responsabilidade é grande. Sem dúvida, o maior desafio será atender e satisfazer todos os *stakeholders* envolvidos no assunto Big Data, de forma a atender aos anseios dos negócios, prover melhores experiências para o indivíduo, mas com a manutenção da sua privacidade. O modelo de negócio baseado em dados também é requisito mandatário para as organizações que desejam ter sucesso nessa área.

## TEORIA EM PRÁTICA

Toda organização precisa estabelecer um modelo de negócio, ou seja, sua forma de atuação na sua indústria. Um modelo de negócios é a forma como uma organização está arranjada ou representada para criar valor na sua indústria de competição. Geralmente, os modelos de negócios são resultados da definição do posicionamento estratégico, considerando a visão, missão, os objetivos e outros elementos. Dada a importância dos dados, eles vêm sendo considerados um elemento influenciador mandatário nos modelos de negócios. As organizações já estabelecidas

no mercado por um longo tempo têm dificuldade em flexibilizar e inovar seus modelos de negócio, a barreira cultural e a equipe madura dificultam a mudança. No entanto, as organizações *startups* conseguem definir modelos de negócios inovadores baseados em dados e encontram poucas barreiras de entrada às indústrias. Exemplos mais conhecidos são: Netflix, Uber e Airbnb. Para superar a barreira da inovação, as organizações mais tradicionais têm procurado disponibilizar para seus colaboradores novas formas de usar e analisar os dados, ações de melhoria de processos, compartilhamento da base de dados. Essas ações visam manter o funcionamento normal dos negócios e aproveitar os dados de forma mais eficaz. Os modelos de negócios baseados em dados irão substituir os modelos tradicionais menos eficientes, pois esses modelos proporcionam a criação de novos valores, descobrimento de novos mercados, formas de engajamento, segmentação de nicho e aumentam a interação com o cliente.

Diante desse cenário, pesquise organizações que criaram modelos de negócios e obtiveram sucesso, faça uma lista.

Escolha alguns modelos de negócios baseados em dados e faça uma análise para descobrir quais são os dados críticos de sucesso desse modelo.

## VERIFICAÇÃO DE LEITURA

1. Atualmente, o assunto Data Science e Big Data estão em boa parte das agendas dos executivos, tão relevante

que o mercado não hesita em afirmar que o dado é tão importante, que compará-lo à importância do petróleo para a sociedade em décadas passadas não é exagero, por isso, dizem que o dado é o petróleo da era digital. Justifica essa afirmação:

- a. O crescimento vertiginoso dos dados com duplicação a cada 18 meses. A conclusão é que temos muitos dados à disposição para análise e geração de negócios.
  - b. O crescimento vertiginoso dos dados com duplicação a cada 18 meses. A conclusão é que esse volume enorme pode ser vendido e comercializado.
  - c. A análise de dados que incentiva a criação de carros elétricos e autônomos que gera desinteresse pelo petróleo.
  - d. O crescimento natural da geração de dados, por vezes até gratuita; em contrapartida, a exploração do petróleo é cara e poluidora.
  - e. O fato de o dado poder ser decomposto e gerar vários produtos como o petróleo.
2. A evolução do Business Inteligente e análise de dados são demonstradas nas afirmativas abaixo. Assinale a alternativa correta.
- I. Evolução do armazenamento de dados passou por bancos de dados estruturados evoluindo para bases não estruturadas com influência da web e tecnologias de mobilidade e sensores.
  - II. Aplicações abrangem áreas, comércio eletrônico, governo, política, pesquisa, saúde e segurança, entre outras.

III. Pesquisa emergente está focada em análise de Big Data, web, texto, rede e mobilidade.

- a. I e II.
- b. II e III.
- c. I e III.
- d. Somente a II.
- e. I, II e III.

3. A organização deve ter a capacidade de criar uma estrutura para lidar com a análise de dados, necessita contar com formas de desenvolver e usar os recursos humanos e técnicos relacionados a Big Data.

#### PORQUE

O mercado oferece inúmeras formas de coletar, acessar, rastrear, gerir, processar, analisar e dispor os dados para a tomada de decisão. A dificuldade não está somente em adquirir ou desenvolver os recursos, mas também em como dispor esses recursos e como estruturar a equipe.

Com base nas informações dadas e na relação proposta entre elas, é CORRETO afirmar que:

- a. As duas afirmações são verdadeiras e a segunda justifica a primeira.
- b. As duas afirmações são verdadeiras e a segunda não justifica a primeira.
- c. A primeira afirmação é verdadeira e a segunda é falsa.
- d. As duas afirmações são falsas.
- e. A primeira afirmação é falsa e a segunda é verdadeira.

## Referências bibliográficas

- O'CONNELL, A.; FRICK, W. **From data to Action.** HBR – Harvard Business Review, 2014. Disponível em: <[https://www.sas.com/content/dam/SAS/en\\_us/doc/whitepaper2/hbr-from-data-to-action-107218.pdf](https://www.sas.com/content/dam/SAS/en_us/doc/whitepaper2/hbr-from-data-to-action-107218.pdf)>. Acesso em: 30 ago. 2019.
- DAVENPORT, T. H. How strategists use “Big Data” to support internal business decisions, discovery and production. **Strategy and Leadership**, v. 42, n. 4, p. 45-50, 2014.
- CHEN, H.; CHIANG, R. H.; STOREY, V. C. Business Intelligence and Analytics: From Big Data to Big Impact. **MIS quarterly**, v. 36, n. 4, p. 1.165-1.188, 2012.
- SILVERIA, M.; MARCOLIN, C. B.; FREITAS, H. M. R. Uso corporativo do Big Data: uma revisão de literatura. **Revista de Gestão de Projetos – GeP**, v. 6, n. 3, set./dez. 2015.
- HUANG, T.; Van Mieghem, J. A. (2014). Clickstream data and inventory management: Model and empirical analysis. **Production and Operations Management**, v. 23, n. 3, p. 333-347.
- GÜNTHER, W. A.; MEHRIZI, M. R.; HUYSMAN, M.; FELDBERG, F. Debating Big Data: A literature review on realizing value from Big Data. **Journal of Strategic Information Systems**, Amsterdã, v. 26, p. 191-209, 2017.

## Gabarito

### **Questão 1 – Resposta A**

**Resolução:** o crescimento do volume e a variedade do dado com a possibilidade de análise e obtenção de insight está permitindo modelos de negócios geradores de caixa para muitas organizações, isso o torna tão valioso.

### **Questão 2 – Resposta E**

**Resolução:** as três afirmações estão corretas e representam a evolução do BI, o surgimento de aplicações em várias áreas e o foco das pesquisas emergentes atualmente.

### **Questão 3 – Resposta A**

**Resolução:** as duas afirmações são verdadeiras e se justificam, elas tratam do conceito e da dificuldade que uma organização tem para definir a estratégia de governança dos dados.



# ***Big Data Analytics***

Autor: Aimar Martins Lopes

## **► Objetivos**

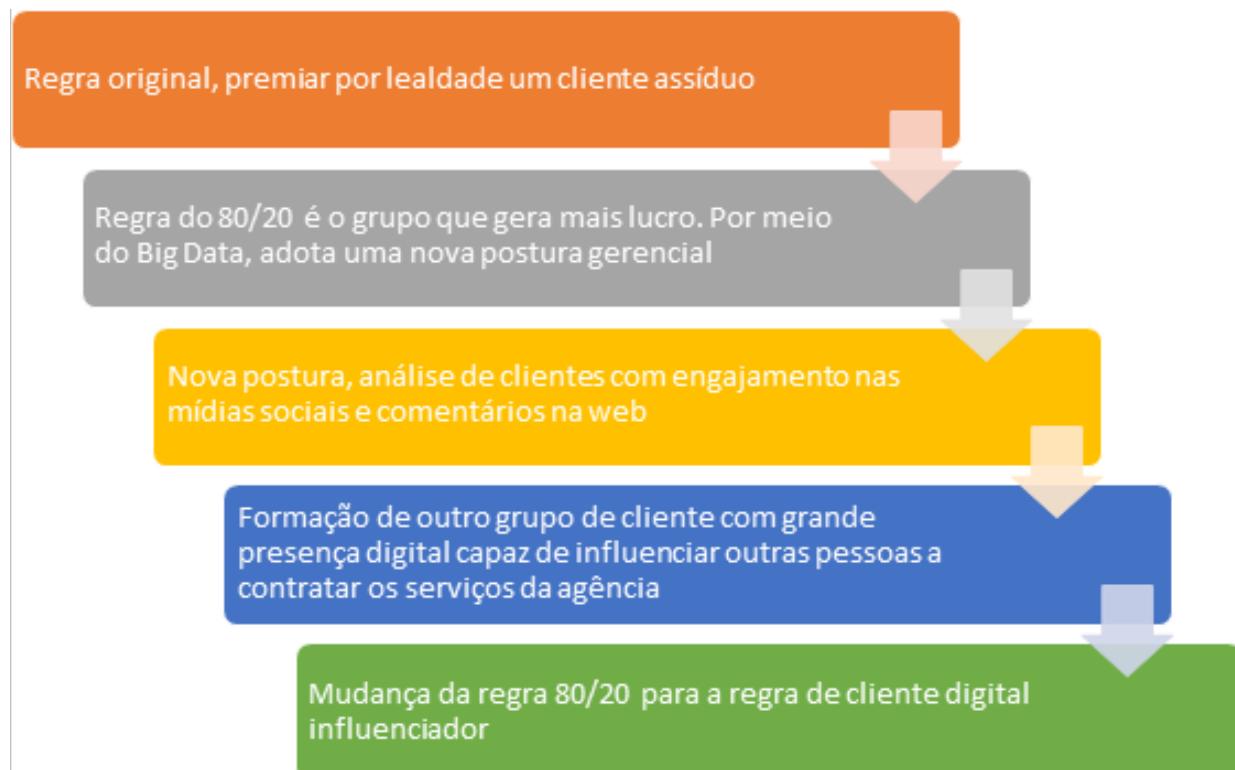
- Conhecer e avaliar as tecnologias do *Big Data*.
- Compreender os impactos do *Big Data*.
- Analisar e propor a inclusão de análise de dados em um negócio.
- Demonstrar e avaliar as potencialidades do *Big Data* na tomada de decisão.

## ► 1. De *Big Data* para *big impacto*

Hábito de longas décadas, a maioria das organizações trabalha arduamente para atender à regra do 80/20 (princípio de Pareto), conhecido também como regra dos poucos vitais. Quais são os 20% dos meus clientes que geram 80% do faturamento? Quais 20% dos produtos são responsáveis por 80% das vendas? Quais os 20% das reclamações do cliente? Quais os 20% dos motivos que geram 80% das reclamações? Mesmo as organizações com *dashboards* de KPI (Key Performance Indicator) têm dificuldade em conhecer qual análise dos 80/20 é a melhor.

Para ilustrar a ideia acima, vamos imaginar uma situação típica de uma agência de viagem. Veja no Quadro 4 a mudança de uma regra de negócio.

Quadro 4 – Impacto do *Big Data* numa agência de viagem



Fonte: elaborado pelo autor

O impacto do *Big Data* numa agência de viagens, conforme o exemplo acima, é comum na maioria das agências. Suas regras são voltadas para beneficiar ou premiar por lealdade um cliente assíduo, portanto, a regra do 80/20 atenderá bem, provavelmente é o grupo que gera mais lucro para o negócio. Contudo, por meio de Big Data e nova postura gerencial, a organização passa a cruzar as informações dos melhores clientes e seu engajamento nas mídias sociais e comentários na web referentes às viagens e aos serviços prestados pela agência.

Essa análise de dados pode resultar na formação de outro grupo de cliente, com grande presença digital, capaz de influenciar outras pessoas a contratar os serviços da agência. Diante disso, a organização pode mudar sua regra de lealdade para a de influenciador, este último passará a usufruir dos privilégios a partir de agora. Esse novo grupo é capaz de aumentar a venda da agência, pois trará um grande número de clientes.

## ► 2. Big Data e seus caminhos

Big Data é um conjunto de tecnologias para lidar com grandes volumes de dados, a análise de dados se refere as tecnologias de *business intelligence* e *analytics* (BI&A), ambas utilizadas em *machine learning* (aprendizado de máquina) e análise estatística. Muitas dessas tecnologias já estão no mercado há um bom tempo, banco de dados relacional (SGBRr – Sistema de Gerenciamento de Banco de Dados Relacional), *data warehouse*, ETL (extrair transformar e carregar), OLAP (Online Analytical Processing), entre outras.

A partir da disseminação dos microcomputadores na década de 1980, houve um crescente desenvolvimento de várias ciências e tecnologias da informação. No início de 1990, vários algoritmos de mineração de dados foram desenvolvidos e aprimorados até os dias atuais.

Os algoritmos de análise de dados indicados como referência são o C4.5, K-Médias, SVM (máquina de vetores de suporte), EM (maximização de expectativa), PageRank, AdaBoost, KNN (vizinhos K-próximos), Naïve Bayes e CART (WU et al., 2007). Esses algoritmos cobrem classificação, agrupamento, regressão, análise de associação, análise de rede e *machine learning*. Atualmente, a maioria desses algoritmos populares de mineração de dados foi incorporada em sistemas de mineração de dados comerciais e de código aberto. Em outra ponta, o crescimento de Big Data é avançado pelo uso das redes neurais, que permitem o aprendizado de máquina (*machine learning*), predição, *clustering* (agrupamento) e desenvolvimento de algoritmos genéticos (CHEN et al., 2012).

Outras abordagens também coexistem no mundo da análise de dados, uma delas é fundamentada em teoria e modelo estatístico. A aplicação da análise multivariada com técnicas de análise fatorial, regressão, discriminante e *clustering* são muito utilizadas nas aplicações de negócios pelas organizações. A outra é fundamentada nas técnicas de heurística e otimização, pois são adequadas à situação comercial em que um limite deve ser imposto.

O aprendizado de máquina envolvendo a estatística, geralmente baseado em modelos matemáticos, é utilizado em aplicativos de dados, texto, rede social e web. Outros algoritmos e técnicas exploram situações exclusivas, tais como: análise sequencial, temporal e espacial, bem como fluxos de dados de sensores e de alta velocidade. Menos comum é a mineração de processo realizada por meio de eventos registrados pelos setores aos quais uma tarefa passa. Exemplo: cadeia de suprimento logístico.

## ASSIMILE

O *machine learning* (aprendizado de máquina – ML) geralmente se refere às mudanças nos sistemas que executam tarefas que simulam o raciocínio humano

(inteligência artificial – IA). Essas tarefas estão relacionadas ao reconhecimento, diagnóstico, planejamento, controle de robô, previsão, etc. As mudanças propostas pelo ML são várias, geralmente são melhorias para sistemas já em execução ou síntese de novos sistemas. O agente da mudança percebe e modela seu ambiente e calcula ações apropriadas, às vezes antecipa seus efeitos. As ações que promovem as alterações nos sistemas são chamadas de aprendizado (NILSSON, 1998).

### ► 3. Big Data e a análise de textos

Grande parte dos dados coletados pela organização é textual (não estruturado), ou seja, está em formato de texto de e-mail, blog, páginas na web, documentos corporativos, mensagens trocadas na mídia social, etc. Mas, sem dúvida, a área mais quente é a da web (internet), tudo pode ser registrado e analisado, um elemento a se destacar está na classificação desses dados, que são conhecidos como não estruturados.

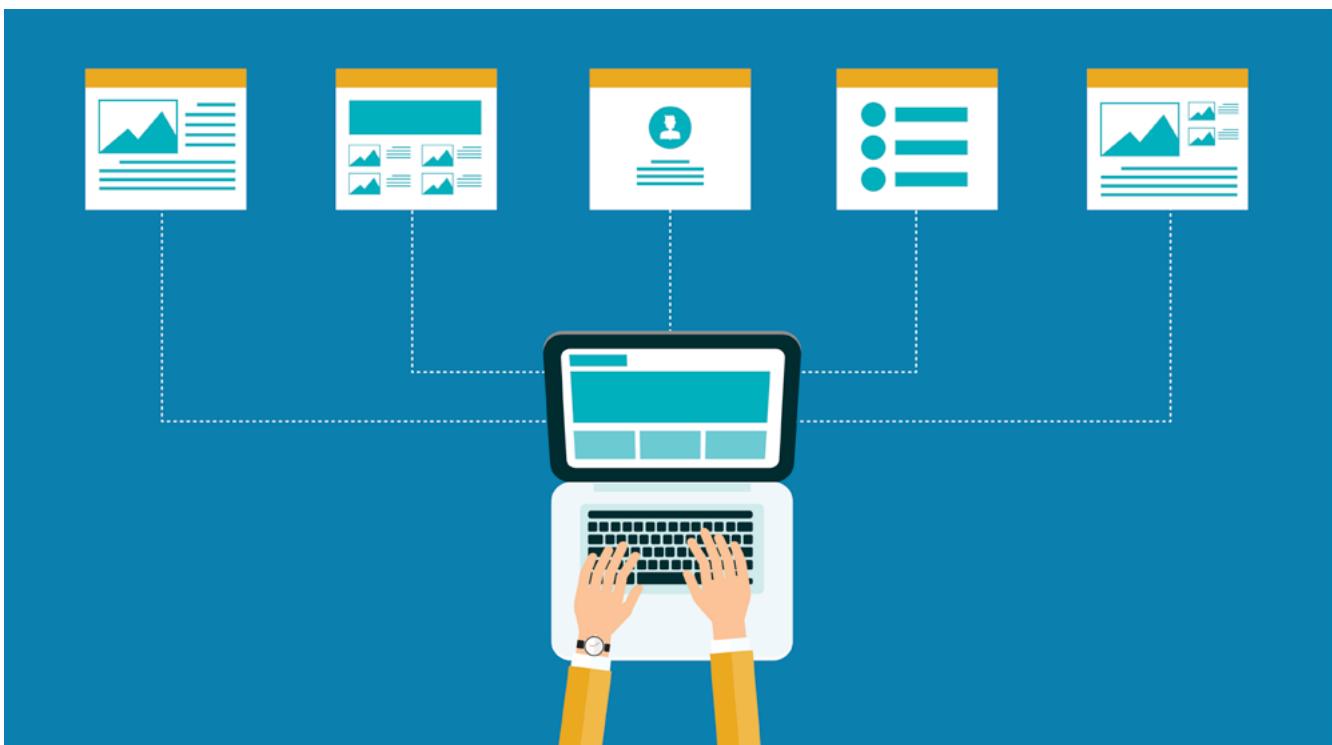
No início do tratamento de texto, as áreas analisadas eram de rastreamento distribuído, análise de registros, ranking de páginas, indexação invertida, todo esse conhecimento foi implantado nos sistemas de gerenciamento de documento e nos algoritmos de BI. Atualmente, as técnicas de análise de texto, extração de informações, modelos de tópicos, questionários (Q/A) e mineração de opinião estão em alta. No campo de extração de informações em texto, o NER (reconhecimento de entidade nomeada), identifica elementos atômicos e os classifica em categorias predefinidas (nome, idade, data de nascimento, etc).

Nos modelos de tópicos, os algoritmos buscam temas em um conjunto de documentos não estruturados, o algoritmo LDA (alocação latente de

Dirichlet) pode ser utilizado para esse fim. Na sequência, os sistemas de resposta a perguntas (Q/A), atendidos por técnicas de PNL (programação de neurolinguística), são capazes de recuperar informações e interagir como humanos. Respondem a perguntas, analisam questões, extração e apresentação de respostas, etc. As organizações têm utilizado essa técnica para interação de site, educação, segurança e saúde.

Outra análise textual é a de sentimentos, identifica a opinião por meio da análise textual que identifica afetos, raiva, carinho e outros estados emocionais. As organizações têm interesse em conhecer as opiniões do público e da sociedade em relação a organização, produtos/serviço, tendência política, eventos sociais, estratégias da organização, resultados de campanha de marketing, etc. A Figura 16 ilustra diversas fontes de texto como fontes de dados.

Figura 16 – Fontes de texto diversas



Fonte: TCmake\_photo / iStock

Em suma, essa técnica extrai, classifica, comprehende e avalia as opiniões dos indivíduos postadas em e-mail, mídia social, notícias on-line, etc

## ► 4. Big Data e a análise na internet

A internet é o principal campo para a mineração, com a evolução dos modelos de recuperação de informações, análise estatística, mineração de dados, análise de texto e outros, provêm desafios e oportunidades analíticas importantes. Os hiperlinks HTTP/HTML, mecanismos de pesquisa da web e os sistemas de diretório para conteúdo da internet contribuíram para o surgimento de tecnologias com a função de rastrear informações, rastrear sites, atualizar e classificar sites, analisar pesquisa e, principalmente, analisar logs de transações de clientes.

Outra forma de prestar serviço em Big Data é com o uso de modelos de programação leves que suportam a organização e notificação de dados e de conteúdo multimídia de diferentes fontes para acessar conteúdo de produtos. Com essa tecnologia, um prestador de serviços da internet pode disponibilizar dados para os desenvolvedores de aplicativos, tais como: catálogos de produtos, histórico de preço, avaliação de clientes, acesso ao site, etc. A empresa Google, por exemplo, libera APIs para o Gmail, agenda, mapas, entre outros, para esse fim.

A análise de dados na web tem um componente em crescimento constituído pelas plataformas e pelos serviços de computação em nuvem (*cloud computing*). Neles estão incluídos software de sistema, plataforma de desenvolvimento, hardware da web, aplicativos e serviços. Essa tecnologia é conhecida como SOA (Service-Oriented Architecture), ou seja, arquitetura orientada a serviços. Nela encontramos serviços de processamento de software, plataforma de desenvolvimento de aplicativos e hardware. Tais serviços provenientes da computação em nuvem são denominados de serviço de software (SaaS), plataforma como serviço (PaaS), infraestrutura como serviço (IaaS). Apesar de não ser tão recente, há poucos fornecedores atuando com alto desempenho, a AWS (Amazon Web Service) é um deles.

## 5. O Big Data e a análise de rede

A análise de dados em rede cresceu rapidamente devido às pesquisas de sociólogos, cientistas da computação e matemáticos. As relações de redes mais comuns são cliques, caminhos, centralidade, laços, entrelinhas, análise estrutural de rede social.

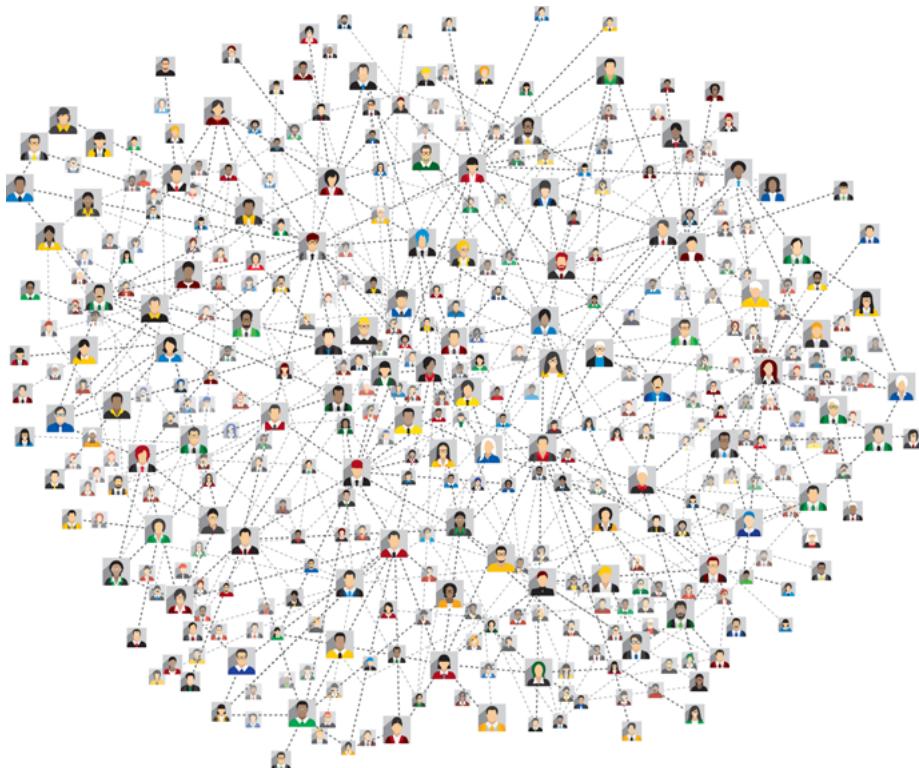
A mineração de links e detecção de comunidades (*cluster*) são exemplos de análise de rede. Nelas procura-se prever ou detectar nós de uma rede. Os nós podem representar um grupo de usuário, comprador ou cliente, produto e serviço. Os links entre os nós representam suas relações, por exemplo, troca de e-mails, compra e venda, colaboração, relação social, etc.

Quando se consegue produzir um grafo de uma rede, os pontos de intensidade também podem representar uma comunidade social. As técnicas podem, inclusive, tratar redes de doenças, comunidades virtuais, terroristas, redes políticas e crime organizado.

A técnica de análise de rede de modelos aleatórios de grafos exponenciais (ERGM) é formada por um grupo de algoritmos estatísticos capaz de analisar dados de redes sociais (relacionamento, amizade, saúde). Eles conseguem entender as propriedades latentes causadoras da formação da rede social, sua evolução e modificação, seja ela uma rede de doença, cliente, aluno, torcida ou paciente.

Para a análise de dados em geral, os cientistas de dados fazem uso intenso de modelo estatístico. Um modelo estatístico pode ser entendido quando há várias hipóteses sobre um conjunto de dados observados e dados semelhantes de diferentes fatores. O modelo deve ser capaz de descrever a relação entre variáveis aleatórias com outras variáveis não aleatórias. Geralmente é representado por um par (OP DP), em que OP é a amostra – conjunto de Observações Possíveis – e DP é um conjunto de Distribuição de Probabilidade.

Figura 17 – Agrupamento de uma comunidade (cluster)



Fonte: Aelitta / iStock (c).

## ► 6. O Big Data e a mobilidade

A cultura da sociedade atual é do “aqui e agora”, então, para atender a essa necessidade, temos que pensar na mobilidade e na computação móvel. De fato, é um canal eficaz para alcançar um enorme público e também é um meio de aumentar a produtividade (eficiência) e o resultado (eficácia) da força de trabalho de uma organização. Algumas organizações de desenvolvimento de software entendem que a computação móvel é a segunda área em demanda de mão de obra e atenção. Nesse sentido, o mobile BI (análise de dados) também é considerado como uma das novas tecnologias com potencial para modificar os negócios.

O marketing para o dispositivo smartphone, cresce à medida que o número de smartphones vendido aumenta. Com ele cresce também a computação móvel e o profissional de TI, pois mais e mais aplicativos são criados.

A Apple App Store, loja de aplicativos da Apple, oferece centenas de milhares de aplicativos em qualquer categoria. A Play Store, loja de aplicativos do Android, também possui quantidade semelhante.

Os *apps* de análise de dados para a área móvel são bem diversificados, mas podemos encontrar aplicativos específicos do negócio, específicos do setor, corporativos, aplicativos de *e-commerce*, gamificação e aplicativos sociais. Num jargão técnico, a programação dos serviços da internet (HTML, XML, CSS, Ajax, Flash, J2E) e as plataformas Android e iOS contribuem para o desenvolvimento de serviços da internet móvel.

O smartphone, por ser um dispositivo inteligente, abre possibilidades inovadoras para Big Data e análise de dados. Dispõe de aplicativos avançados e inovadores capazes de coletar conteúdo personalizado, que retrata um contexto real. Como dispõe de hardware e conteúdo, a plataforma de aplicativos móvel pode reunir uma comunidade inteira de desenvolvedores voluntários, isso, sem dúvida, oferece um novo caminho para a pesquisa de análise de dados. A Figura 18 representa dados à disposição do toque.

Figura 18 – Mobilidade e toque



Fonte: Akindo / iStock

A análise de dados móvel tem muito a crescer, pois tecnologia surge em diversas áreas e rapidamente se torna popular. Algumas são: aplicativos de detecção de dispositivos móveis sensíveis à atividade e localização, inovação social móvel para *m-education*, *m-learning* e *m-health*, *crowdsourcing*, visualização, redes sociais móveis, personalização, modelagem comportamental e sentimental, gamificação, marketing e publicidade social, etc.

## ► 7. Big Data e a criação de valor

As organizações mais espertas já descobriram a importância de obter informação, contextualizar e criar *insight* com alto potencial de valor. Os executivos do alto escalão querem saber se estão absorvendo todos os valores dos dados.

As organizações com melhor desempenho usam a análise de dados cinco vezes mais do que outras organizações de menor desempenho. Seus estudos mostraram que a análise oferece valor aos negócios. Os executivos afirmam que a melhoria da informação e da análise é uma das principais prioridades das suas organizações. E mais de um em cada cinco disse estar sob pressão intensa ou significativa para adotar abordagens avançadas de informação e análise (LAVALLE et al., 2011).

Os executivos de alto escalão querem que as organizações executem decisões baseadas na análise de dados, desenvolvam cenários, simulações para decisões imediatas e situações de interrupções, exemplo: surgimento de correntes inesperados ou um terremoto em uma indústria de suprimento, cliente sinalizando o desejo de mudar de fornecedor, etc. Os executivos querem entender as soluções ideais com base em parâmetros de negócios complexos ou novas informações, e querem agir rapidamente (LAVALLE et al., 2011).



## PARA SABER MAIS

Atenção, é preciso saber quem valorizaria o que, por que e o quanto ele é importante, e às vezes ele não é o seu cliente-alvo tradicional. Procure identificar os clientes-alvo lançando contato com uma ampla rede de relacionamento, realize entrevistas com clientes para entender seus negócios e processos, procure lacunas de dados e identifique como preenchê-las por meio de dados, produtos, serviços, etc. Exemplo: a percepção positiva da classe médica sobre um medicamento pode ser positiva para gestores de ativos financeiros, pois indica oportunidade de investimento no laboratório fabricante.

Adotar uma metodologia de cinco pontos para implementar com sucesso o gerenciamento orientado por análise de dados e gerar valor rapidamente é o que recomenda Lavalle et al. (2011).

1. Primeiro, pense grande: foco nas oportunidades maiores e mais valiosas.

Atacar o maior desafio é arriscado, contudo, quando as apostas de um projeto são grandes, a alta gerência é ativa e os melhores talentos buscam se envolver. Por outro lado, não comece a fazer análises sem a orientação estratégica dos negócios.

2. Comece no meio: escolha e comece com perguntas, não com dados.

As organizações devem começar pelo meio do processo, implementando a análise e definindo primeiro os *insights*, depois as perguntas necessárias para atender ao objetivo da organização e, em seguida, identificar os resultados.

3. Manter o Analytics vivo: incorpore *insights* para impulsionar ações e agregar valor.

Incorpore novos métodos e ferramentas para adicionar informações aos processos de negócios – casos de uso, soluções analíticas, otimização, fluxos de trabalho e simulações estão tornando as percepções mais compreensíveis e açãoáveis, como também a análise de tendências, previsão e os relatórios padronizados, modelagem e visualização.

4. Adicione, não se desvie: mantenha os recursos existentes e adicione novos recursos.

À medida que os recursos analíticos são adicionados no início dos níveis de gerenciamento centrais, os recursos existentes não devem ser subtraídos. À medida que novos recursos são incorporados, os existentes devem continuar sendo suportados.

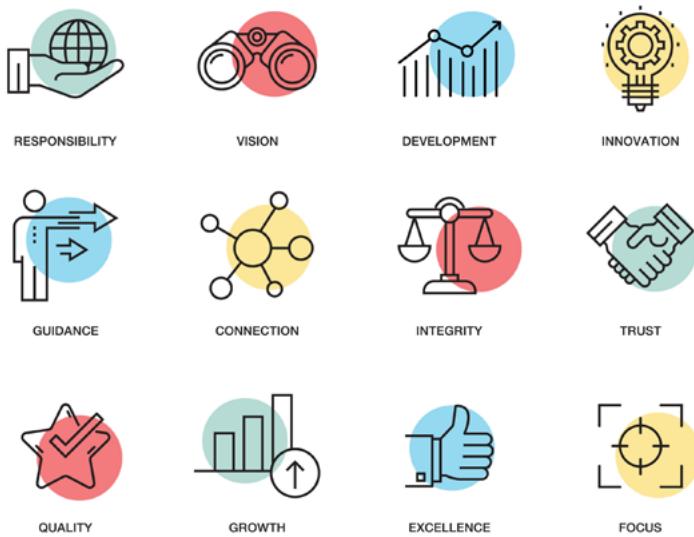
5. Construa as peças, planeje o todo: use uma agenda de informações para planejar o futuro.

Informações estratégicas começaram a chegar por meio de canais digitais não estruturados: mídias sociais, aplicativos de smartphones e um fluxo cada vez maior de dispositivos emergentes baseados na internet. Todos esses dados devem ser moldados em uma base de informações integrada, consistente e confiável, portanto, mesmo que as organizações iniciem o caminho da análise atacando seletivamente os maiores problemas.

Para gerar valor para os clientes por meio de modelos de negócios baseados em análise de dados, mantenha o alinhamento da estratégia com a área de dados, trabalhe para que toda a equipe tenha foco nos grandes problemas, selecione os desafios capazes de serem resolvidos e planeje uma agenda para o futuro.

Figura 19 – Valores centrais

## CORE VALUES COLOR SPOT ICONS



Fonte: Cnythzl / iStock

## ► 8. Big Data Analytics e o processo de decisão

A decisão de um indivíduo pode ser induzida ao erro em muitos casos, especialmente quando há muita informação envolvida, mesmo que se considere a experiência e intuição no processo de tomada de decisão. Um bom método de tomada de decisão integra componentes que garantem a qualidade e a relevância de um sistema de apoio à decisão, bem como os dados analisados. Um método recomendado se inicia pelo processo de preparação, análise de dados e tomada de decisões. Utilize também os *insights* gerados ao longo do tempo e outras informações armazenadas em um banco de dados de conhecimento. A Figura 20 mostra um *framework* de tomada de decisão.

Figura 20 – Método para tomada de decisão



Fonte: elaborado pelo autor.

A eficiência e eficácia do método depende também da integração do sistema do fluxo de análise de dados com o sistema de armazenamento de dados, ambos vinculados ao sistema executivo de tomada de decisão. Para compreender melhor, vamos descrever os componentes do *framework*:

## 1. Preparação de dados

O sistema de suporte à decisão tem interface com outros sistemas de dentro da organização ou de fora por meio de serviços e interfaces da web. As fontes de dados são provenientes de fontes de dados tradicionais, como bancos de dados ou sistemas transacionais e informacionais, fluxos de dados, mídias sociais, IoT e dispositivos móveis. A preparação segue a ordem:

- Filtros de qualidade de dados atuam em grande escala, estruturando os dados por fatores de qualidade, relevância, precisão, metadados, disponibilidade e integridade;

- Integração de dados relevantes decide quais fontes de dados serão integradas e filtradas;
- Preparação de dados para análise transforma e limpa os dados, deixando-os prontos para análise.

Ao final, os dados que serão analisados devem ter qualidade e relevância, além de possibilidade de juncão com dados de contexto anterior.

## 2. Análise de dados

O objetivo da análise de dados é gerar sugestões para a decisão baseadas em modelo descritivo ou preditivo, em um algoritmo e fatores externos relevantes para o contexto e o problema investigado. Seguem os passos:

- Recebe os dados filtrados e preparados da preparação dos dados;
- Seleciona um algoritmo apropriado com base em fatores relevantes ao problema. O algoritmo deve estar relacionado ao negócios-alvo da organização;
- Executa a análise analítica preditiva necessária;
- Cria um modelo de análise preditiva com resultados potenciais;
- Apresenta ao tomador de decisão o resultado como forma de recomendação e visões (*insight*).

## 3. Tomada de decisão

O indivíduo tomador de decisão recebe um conjunto de sugestões e visões, acrescenta suas experiências, interage com as ferramentas de visualização e seleciona os elementos relevantes, encontra alternativas e, então, seleciona a melhor decisão e os planos para os objetivos de

negócios. Propõe também mudanças de posicionamento, propósito, objetivos, serviços, produtos, entre outras.

Todos os elementos dispostos ao tomador de decisão devem garantir possibilidade de exploração e alternativas dos fluxos de dados filtrados, fontes tradicionais, além da recomendação preditiva.

#### **4. Insights**

O *insight* é composto por ações que registram as decisões em repositórios para fornecer *feedback* em ciclos do método de tomada de decisão, ou seja, preparação de dados, análise de dados e tomada de decisões. O processo de *insight* é vinculado a uma base de conhecimento capaz de classificar, filtrar e organizar *insights*. As organizações mais evoluídas possuem sistema automatizado para melhorar o desempenho dos ciclos de decisões e alertar em casos de ineficiências das decisões.

Finalizando, no Quadro 5, sugerimos algumas tendências no campo Big Data Analytics.

Quadro 5 – Algumas tendências em Big Data Analytics

TENDÊNCIAS
<ul style="list-style-type: none"><li>• Organizações baseadas em informações</li><li>• Sistemas vão substituir os processos baseados em humanos</li><li>• Automação ampla de processos e máquinas</li><li>• Análise de Big Data em toda a organização</li><li>• Evolução das técnicas avançadas de Big Data</li><li>• Crescimento contínuo da IoT</li><li>• Uso para proteção de ameaças à segurança cibernética</li><li>• Mudança nos modelos de operações</li><li>• Intensidade na criação de oportunidades e modelos de negócios e sociedade</li></ul>

Fonte: elaborado pelo autor.

Diante dos temas discutidos a respeito do Big Data Analytics, pode-se compreender melhor a influência dos dados nos negócios e na tomada de decisão como também a influência da internet. As implicações do impacto dessa tecnologia abrange análise de texto, redes, web e mobilidade, mostra a diversidade e a potencialidade de compreensão dos novos modelos de negócios. Os métodos apresentados, com certeza, mudam radicalmente a forma de modelar os negócios. Uma organização que deseja se perpetuar deve, urgentemente, se já não está, incluir na sua agenda a missão de criar uma estrutura de Big Data Analytics, bem como a governança dos dados.

## TEORIA EM PRÁTICA

### **Jaguar Land Rover Automotive rumo à 4<sup>a</sup> Revolução Industrial**

A maior fabricante de automóveis do Reino Unido, a Jaguar Land Rover Automotive, apresenta reputação de inovação e por isso caminha para o futuro por meio de investimentos em soluções e aplicativos com tecnologias de Big Data e algoritmos de aprendizado.

Em evento realizado pela fabricante em 2017, foram abordadas discussões de como viver em um mundo mais conectado, exposições e veículos elétricos que permitem manter o prazer de dirigir em carro autônomo. No mesmo evento, foi apresentado o Sayer, um volante inteligente ativado por voz. O volante Sayer revoluciona a maneira como se dirige, além de se tornar um dispositivo móvel conectado com assistente de voz. Os motoristas do futuro levarão seu volante com eles e esta poderá ser a única parte do carro que possuem. O volante Sayer também será o “cartão de associação” para um clube de serviços sob demanda. Com recursos semelhantes aos de um assistente

virtual, como o Alexa da Amazon, o volante poderá fazer reservas, acessar sua agenda e ler as notícias para você.

Outra inovação, associada às barreiras psicológicas que se referem à confiança em carros autônomos, está na inserção de olhos no carro autônomo para se comunicar com os pedestres. Os olhos têm a função de conexão com um pedestre da mesma maneira que um motorista humano faria um contato visual para permitir que o pedestre saiba que ele o vê e parará para permitir que ele cruze uma rua com segurança.

(Fonte: MARR, 2018)

A ações da Jaguar Land Rover Automotive são um exemplo da evolução e adoção das tecnologias digitais na indústria.

Pesquise e aprenda com outros casos semelhantes. Escolha alguns exemplos e faça uma lista das tecnologias e ações adotadas com relação a Big Data Analytics.

## VERIFICAÇÃO DE LEITURA

1. Hábito de longas décadas, a maioria das organizações trabalha arduamente para responder ao princípio de Pareto – regra do 80/20, conhecido também como regra dos poucos vitais. Quais são os 20% dos meus clientes que geram 80% do faturamento? Quais 20% dos produtos são responsáveis por 80% das vendas? Quais os 20% das reclamações do cliente? Quais os 20% dos motivos geram 80% das reclamações? Mesmo as organizações de alta performance, repletas de *dashboards* de KPI (Key Performance Indicator),

têm dificuldade em encontrar qual análise dos 80/20 é a melhor. De acordo com o texto acima, NÃO podemos afirmar:

- a. A regra de Pareto 80/20, apesar de ser muito útil, não é mais utilizada pelas organizações.
  - b. A regra de Pareto 80/20 indica que 20% de algo representa 80% de alguma coisa, ou seja, 20% dos clientes geram 80% do faturamento.
  - c. Um *dashboards de KPI* (Key Performance Indicator) é uma boa ferramenta de controle.
  - d. O crescimento do Big Data Analytics irá fazer com que as empresas deixem de utilizar o princípio de Pareto 80/20.
  - e. Ainda hoje, a maioria das organizações trabalha arduamente para responder ao princípio de Pareto – regra do 80/20.
2. As afirmações abaixo se referem ao Big Data e à análise de rede. Leia e assinale a alternativa correta.
- I. As relações de redes mais comuns são cliques, caminhos, centralidade, laços, entrelinhas, entre outros.
  - II. Os links entre nós de redes representam as relações desses nós, por exemplo, troca de e-mails, compra e venda, colaboração, relação social, etc.
  - III. A técnica de análise de rede de modelos aleatórios de gráficos exponenciais é formada por um grupo de algoritmos estatísticos capaz de analisar dados de redes sociais (relacionamento, amizade, saúde).
- a. I e II.
  - b. II e III.

- c. I e III.
  - d. Somente a II.
  - e. I, II e III.
3. Um indivíduo pode tomar uma decisão errada quando há muitos dados envolvidos, mesmo que considere sua experiência e intuição.

#### PORQUE

Um bom método de tomada de decisão integra componentes de qualidade relevantes para auxiliar no apoio à decisão.

Com base nas informações dadas e na relação proposta entre elas, é CORRETO afirmar que:

- a. As duas afirmações são verdadeiras e a segunda justifica a primeira.
- b. As duas afirmações são verdadeiras e a segunda não justifica a primeira.
- c. A primeira afirmação é verdadeira e a segunda é falsa.
- d. A primeira afirmação é falsa e a segunda é verdadeira.
- e. A primeira afirmação é falsa e a segunda é falsa.

## ► Referências bibliográficas

CHEN, H.; CHIANG, R. H.; STOREY, V. C. Business Intelligence and Analytics: From Big Data to Big Impact. **MIS quarterly**, v. 36, n. 4, p. 1.165-1.188, 2012.

DAVENPORT, T. H. How strategists use “Big Data” to support internal business decisions, discovery and production. **Strategy and Leadership**, v. 42, n. 4, p. 45-50, 2014.

LAVALLE, S.; LESSER, E.; SHOCKLEY, R.; HOPKINS, M. S.; KRUSCHWITZ, N. Big Data, analytics and the path from insights to value. **Analytics-magazine.org**, 2011. Disponível em: <http://analytics-magazine.org/big-data-analytics-and-the-path-from-insights-to-value/>. Acesso em: 5 abr. 2019.

MARR, B. How Jaguar Land Rover Is Getting Ready For The 4th Industrial Revolution: AI & Autonomous Cars. **Forbes**, 2018. Disponível em: <https://www.forbes.com/sites/bernardmarr/2018/10/26/how-jaguar-land-rover-is-getting-ready-for-the-4th-industrial-revolution-autonomous-vehicles/#6a19deba3a5e>. Acesso em: 9 abr. 2019.

NILSSON, N. J. **Introduction to machine learning**. Robotics Laboratory Department of Computer Science Stanford University Stanford, CA. 1998. Disponível em: <<https://ai.stanford.edu/~nilsson/MLBOOK.pdf>>. Acesso em: 30 ago. 2019.

## **Gabarito**

### **Questão 1 – Resposta D**

**Resolução:** apesar do crescimento do Big Data Analytics e outras tecnologias, não podemos afirmar que as empresas deixaram de utilizar o princípio de Pareto 80/20 como ferramenta. É uma ferramenta útil que tem seu valor.

### **Questão 2 – Resposta E**

**Resolução:** as três afirmações estão corretas e representam as características do Big Data e a análise de rede conforme apresentado na teoria.

### **Questão 3 – Resposta A**

**Resolução:** As duas afirmações são verdadeiras e se justificam, não há garantia que a tomada de decisão seja 100% correta, os dados possuem vários comportamentos, mesmo quando analisados por um bom processo.



# **Algoritmos de aprendizado de máquina para minerar os dados**

Autor: Aimar Martins Lopes

## **► Objetivos**

- Compreender o que são algoritmos e aprendizado de máquina.
- Capacidade para identificar abordagem de aprendizado de máquina.
- Conhecer a avaliar o uso dos principais algoritmos de aprendizado de máquina.

## 1. O que são algoritmos

A palavra algoritmo não é utilizada em conversas comuns, mas na ciência da computação é universalmente conhecida e utilizada para descrever a maneira adequada de resolver problemas por meio da implementação em programas de computador. A maioria dos algoritmos, com base nas suas estruturas de dados, tem como objetivo organizar os dados envolvidos na ciência da computação.

Quando um programa de computador é desenvolvido, torna-se necessário um grande esforço para entender e mapear o problema que se quer resolver, é aí que entra o algoritmo, por meio da lógica de programação. Ele domina a complexidade e decompõe o problema em tarefas menores com o objetivo de simplificar o algoritmo final.

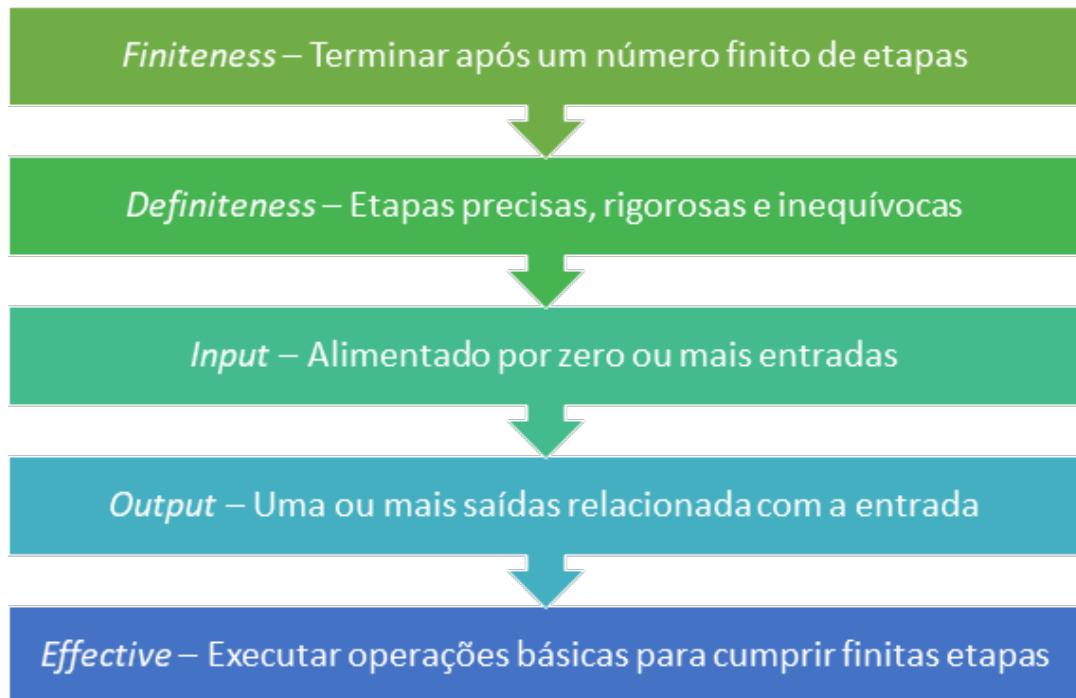
Desenvolver um novo algoritmo ou escolher algum que já esteja pronto para utilizar em um problema é um desafio complexo, envolve muitas variáveis e análise estatística e matemática. Muitos algoritmos demonstram bom desempenho comprovado por análise técnica, porém outros funcionam bem considerando o contexto do problema (SEDGEWICK, 1983).

Um algoritmo eficaz deve:

- Ser um procedimento que reduz problemas a um conjunto de etapas de memorização a serem seguidas;
- Dar uma resposta certa e nunca uma resposta errada;
- Ter um número finito de etapas;
- Trabalhar diversos tipos de problemas.

Além da eficácia, os algoritmos também possuem propriedades comuns entre eles. O Quadro 6 apresenta essas propriedades.

## Quadro 6 – Propriedades comuns dos algoritmos



Fonte: elaborado pelo autor.

É comum pensarmos que os algoritmos são expressados somente pelos programas, mas eles podem ser representados também por fluxogramas ou modelagem de processos que padronizam a compreensão, ou pela linguagem verbal e natural do ser humano carregada de ambiguidade, por pseudocódigo e, finalmente, linguagem de programação que detalha a ação ao nível que o computador comprehende.

### ASSIMILE

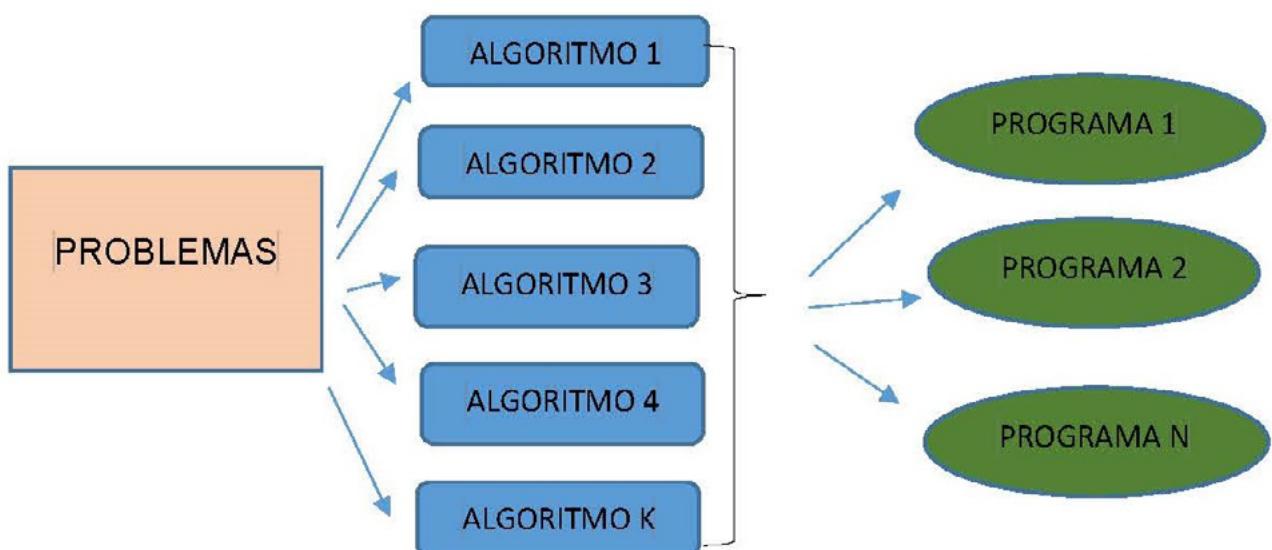
Os algoritmos são interessantes porque eles sempre estão se renovando e novos surgem a todo instante, contudo há alguns que são conhecidos há milhares de anos. No entanto, poucos algoritmos são perfeitamente entendidos. Existem algoritmos complicados, difíceis e complicados, elegantes, simples e fáceis (SEDGEWICK, 1983).

Além de propriedades, o algoritmo também possui uma estrutura composta de:

- *Input*: representa valores de entrada que dimensionam o problema.
- Processamento: são as regras de cálculos matemáticos, aritméticos, analogia, comparações, teste de lógica, etc.
- Seleção: forma de escolher entre dois ou mais cursos de decisão.
- Iteração: processamento repetido e finito de uma coleção de instruções, o término é determinado por uma condição lógica.
- *Output*: apresentar os resultados.

Para compreender a relação dos problemas com os algoritmos, apresento a Figura 21.

Figura 21 – Relação problema x algoritmo x programas traz uma representação da existência do algoritmo



Fonte: elaborado pelo autor

## ► 2. O que é *machine learning* (aprendizado de máquina)

Vamos procurar entender o aprendizado das máquinas (*machine learning*). Isso pode ser facilitado com a ideia do aprendizado animal, pois há vários paralelos entre o aprendizado de máquinas e animais. A aprendizagem está relacionada com uma gama ampla de processos e o conceito de inteligência que torna difícil sua definição. Encontramos sua definição relacionada com aquisição de conhecimento, habilidade em algo, compreensão, instrução ou comando, estudo e experiência. Há também associação com modificação de comportamento em virtude de uma experiência vivenciada. Certamente, os modelos computacionais de aprendizado de máquina são provenientes das teorias de aprendizagem animal.

Se o ser humano pode mudar seu comportamento baseado numa experiência, podemos dizer que o aprendizado de máquina muda o processamento, a estrutura ou programa com base nas relações de suas entradas e saídas externas. Por exemplo, o Google Assistente colhe amostras da sua fala e, a partir desse momento, torna-se capaz de identificar você para uma interação, apesar de ser um exemplo simples, podemos dizer que a máquina aprendeu.

O esforço em fazer as máquinas processarem os dados de forma semelhante à inteligência humana: a inteligência artificial está associada ao aprendizado de máquina por meio de processamento dos dados e execução de tarefas baseados por ações de reconhecimento, diagnóstico, planejamento, controle (automação), predição, entre outras.

O aprendizado de máquina deve se preocupar com o conteúdo da estrutura computacional a ser aprendido. Esse conteúdo é destinado à resolução de algum problema. Para Nilsson, devemos considerar as seguintes estruturas computacionais apresentadas no Quadro 7 (NILSSON, 1998).

Quadro 7 – Estruturas computacionais para *machine learning*



Fonte: adaptado de Nilsson (1998).

## ► 3. Categoria de aprendizado de máquina

O aprendizado de máquina recebe abordagens diferentes com o intuito de melhorar o desempenho dos modelos preditivos. Elas variam de acordo com o problema de negócios envolvido. Vamos ver como são essas abordagens.

### 3.1. Aprendizado supervisionado

Essa abordagem tem o objetivo de encontrar padrões em um grupo de dados cuja classificação é razoavelmente conhecida e, então, aplicar o resultado (padrão encontrado) em um processo analítico. Nesse processo, os dados apresentam seus significados e determinam os rótulos de seus recursos.

Veja o exemplo de uma coleção de imagens de animais. Vamos supor que queremos classificar a coleção em dois grupos por meio do recurso chifre, grupo com chifre e sem chifre. Cria-se, então, um modelo de aprendizado que, após o seu treinamento, considerando o significado

ter chifre (recurso), terá capacidade de analisar fotos novas de animais e classificar o animal no grupo com chifre ou sem chifre.

Para aprimorar o modelo, podemos ir incluindo outros significados e recursos no decorrer do processamento de aprendizagem, de tal maneira que o modelo aprimore a classificação. Isso é possível, pois a regressão usada para o aprendizado colabora para entender a correlação entre as variáveis. Todo o processo envolvido na análise de dados ocorre sem a intervenção do analista de dados. Essa abordagem é muito utilizada no auxílio de problemas de negócios, tais como: análise de seguro, recomendação, análise de risco e crédito, previsão do tempo, etc.

### **3.2. Aprendizado NÃO supervisionado**

É indicado quando o problema analisado possui enorme variedade e quantidade de dados não estruturados ou sem rótulo. Encontramos essa abordagem para análise em dados de mídia social que utilizam a internet como meio de comunicação, exemplo: Facebook, Instagram, Pinterest, Twitter, Snapchat e outros.

O entendimento da estrutura de relação dos dados requer do algoritmo capacidade de classificar o *cluster* (grupo) e o padrão encontrado. São indicados para entender grandes volumes de novos dados não estruturados. A diferença deste para o aprendizado supervisionado é que aqui os dados ainda não são compreendidos. O perfil de clientes que compram em um supermercado é um exemplo de seu uso. Nesse caso, temos a busca de elementos semelhantes entre os clientes que serão agrupados e não a busca de uma variável específica.

Essa abordagem adiciona rótulos aos dados para, em um momento posterior, passarem por uma abordagem supervisionada. Geralmente, o aprendizado não supervisionado é usado como a primeira etapa antes de uma abordagem de aprendizado supervisionado.

Outro exemplo dessa abordagem está na área da saúde; queremos saber se a doença que acomete a população é dengue: com a coleta de grandes quantidades de dados específicos dos pacientes, podem-se auxiliar profissionais da área da saúde a fornecer *insights* relacionando os padrões de sintomas teóricos com os resultados dos dados coletados dos pacientes, identificando quem teve dengue.

### **3.3. Aprendizado por reforço**

Essa abordagem segue a técnica de aprendizado comportamental, por meio de *feedback* da análise dos dados, e direciona o usuário para o melhor desempenho. Nesse aprendizado, o algoritmo não é treinado com uma coleção de dados, o aprendizado se dá por tentativa e erro. Portanto, as decisões de sucesso resultarão no reforço do modelo, pois, aparentemente, ele soluciona melhor o problema.

Essa técnica é muito utilizada para treinar robô. Nesse problema, o algoritmo considera o resultado de suas ações. Um robô de limpeza domiciliar mapeia o ambiente a ser limpado à medida que bate nas paredes: a cada batida, os dados são calibrados para que a navegação seja melhorada, ou seja, o algoritmo é treinado por tentativa e erro para entender o ambiente externo, daí o nome por reforço. Para o sucesso dessa abordagem, o algoritmo deve descobrir a associação do seu objetivo e mapear as dimensões do ambiente, considerando as sequências de suas ações, ou seja, cada trombada em um obstáculo é uma ação de aprendizado.

Devido à complexidade dos carros autônomos, essa abordagem também é utilizada nas ideias que envolvem automação. No mundo animal, podemos fazer a analogia da abordagem do reforço com a técnica de motivação que recompensa o bom comportamento e pune o mau comportamento de um animal. Uma foca, quando

realiza a apresentação solicitada pelo treinador, ganha um peixe de recompensa, portanto, aprende que se repetir a ação receberá novamente um prêmio.

### **3.4. Redes neurais e aprendizagem profunda (deep learning)**

A aprendizagem profunda utiliza a técnica de iteração sucessiva com os dados para a aprendizagem. É profunda, pois incorpora a técnica de rede neural em vários níveis (camadas) sucessivos para aprender. É especialmente recomendada para aprender padrões de dados não estruturados.

Uma rede neural tenta simular o funcionamento do cérebro humano com base no treinamento para solução de problemas mais complexos e pouco definidos. A abordagem de rede neural e aprendizado profundo são recomendados para reconhecimento de voz, imagem, sentimento, comportamento, IoT, rastreamento, etc. A estrutura dessa abordagem é formada pela rede neural composta de camada de entrada, camadas internas (ocultas), onde ocorrem vários níveis de processamento interativo (aprendizado profundo), e camada de saída. Esse processamento recebe o nome de nó (neurônio artificial na rede). Uma rede pode ter até milhões de nós de processamento fortemente interconectados.

Em resumo, temos que *deep learning* (aprendizagem profunda) é parte do constructo *machine learning* (aprendizado de máquina) que faz uso das redes neurais hierárquicas para aprender por meio do uso de algoritmos supervisionados e não supervisionados.

O Quadro 8 abaixo apresenta as quatro abordagens de aprendizado de máquina com as principais características citadas acima.

Quadro 8 – Abordagem de aprendizado de máquina



Fonte: elaborado pelo autor.

## ► 4. Algoritmos de *machine learning*

Uma boa análise de dados e uma decisão acertada sobre o seu resultado deve considerar a escolha do algoritmo mais adequado para o problema a ser analisado como também o modelo de análise escolhido. Um problema pode ser abordado de forma diferente pelos cientistas de dados, por isso que os tipos de aprendizado de máquina servem para ajudar na escolha do melhor. Apesar de diferentes abordagens e construções por diversas organizações e cientistas de dados, uma relação de alguns desses algoritmos, os mais referenciados para o

aprendizado de máquina, será apresentada a seguir. A Figura 22 mostra alguns elementos relacionados a *machine learning*.

Figura 22 – Elementos de *machine learning*



Fonte: PlargueDoctor/iStock.com

## PARA SABER MAIS

O aprendizado de máquina é usado para automatizar o processo de criação de mecanismo de pesquisa na internet, seguem exemplos:

- a. As páginas da web mais relevantes são identificadas pela consulta e estrutura de links da página, conteúdo, frequência de cliques dos usuários e exemplos de consultas de páginas classificadas manualmente.
- b. O aplicativo de filtragem colaborativa das livrarias da internet (Amazon), sites de aluguel de vídeos (Netflix), usam as informações extensivamente para influenciar os usuários a comprar produtos e serviços adicionais (SMOLA, 2010).

**Bayesiano** – Algoritmos bayesianos codificam crenças anteriores sobre como os modelos devem ser. São utilizados quando a quantidade de dados

não é significativa para treinar o aprendizado de máquina e formar um modelo confiável. Seu uso faz sentido quando se conhece previamente parte do modelo. Por exemplo, se conhecemos como um distúrbio cardíaco se apresenta, ou seja, seu modelo ou padrão, quando verificamos o resultado de um diagnóstico de imagens, o algoritmo irá procurar padrões semelhantes ao conhecido nos dados da imagem do paciente.

**Baseado em instância** – Este algoritmo serve para treinar dados, ou seja, monta categoria de novos dados com base na semelhança dos dados de treinamento. Esse conjunto de algoritmo é conhecido como aprendiz preguiçoso, pois não há etapa de treinamento. Ele é formado pela combinação de semelhança pela distância de novos dados com dados de treinamento. É muito útil para reconhecer um padrão, mas não é indicado para conjunto de dados de variação aleatória, dados com valores omissos ou irrelevantes. Um dos algoritmos mais conhecidos é o KNN (K-Nearest Neighbors).

Um exemplo de fácil entendimento é o do torcedor que deseja saber se o clima para o jogo do final de semana será bom. Na tabela de dados, cada linha é uma instância; a primeira coluna é o clima (sol, chuva, nublado), seguido dos atributos temperatura, vento e umidade. Por fim, temos outra coluna que representa a classe (vai ao jogo ou não). O algoritmo fará uma série de comparações com várias etapas para descobrir qual a probabilidade de o torcedor ir ao jogo.

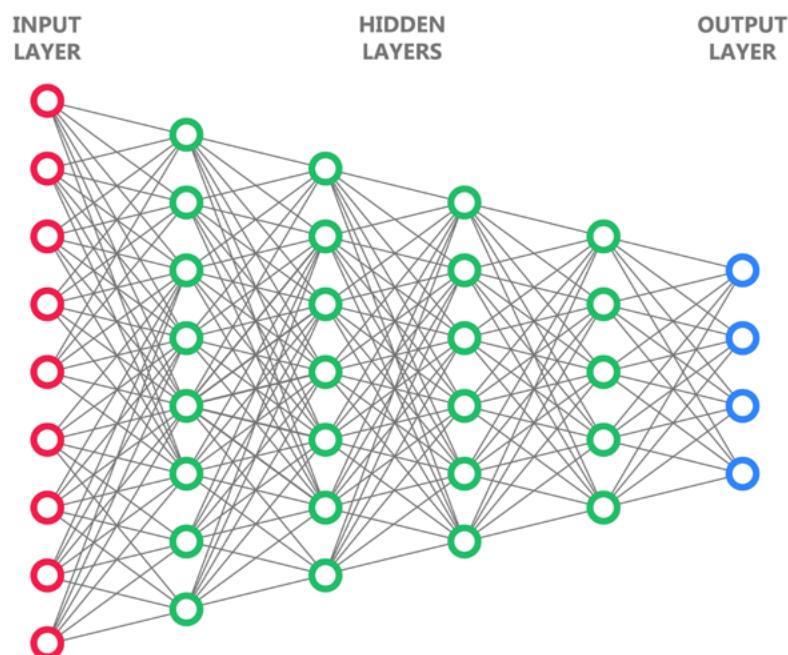
**Clustering** – É uma técnica simples aplicada ao algoritmo que busca entender classificar e agrupar um objeto com características semelhantes. No *clustering*, os objetos agrupados em um cluster são mais semelhantes entre si do que os elementos agrupados em outro cluster. Como os dados não são rotulados, o aprendizado é nomeado de não supervisionado; para agrupar os objetos, o algoritmo interpreta os parâmetros de cada item e agrupa-os. Por exemplo, num rebanho de gado, podemos selecionar os objetos do cluster por meio do agrupamento do gado, considerando o peso, tamanho, idade, sexo, etc.

**Redes neurais** – A principal característica desse algoritmo é a tentativa de imitar o cérebro humano. Com o uso de camadas compostas e interconectadas, aborda os problemas dos dados com base na tentativa de aprender e inferir nos relacionamentos desses dados.

Esse algoritmo é capaz de aprender com a modificação dos dados, por isso é indicado para análise de dados não estruturados ou sem *label* (rótulos). Dada essa característica, o algoritmo é frequentemente utilizado em diversas aplicações, desde análise de diagnóstico médico até carros autônomos. Neste último, a rede de análise fornece a compreensão do ambiente em volta do carro por meio da captura de imagens e sensores.

**Aprendizado profundo (*deep learning*)** – Tem como objetivo ensinar as máquinas a pensar como os humanos. Leva esse nome pois os modelos criados com essa técnica são treinados com um grande conjunto de dados com *label* (rótulo) e redes neurais de várias camadas. Sua evidência está atribuída à sua precisão e, em alguns casos, à superação do desempenho humano.

Figura 23 – Rede neural de vários níveis



Fonte: all\_is\_magic / iStock

**Árvore de decisão** – Os algoritmos de árvore de decisão levam esse nome pois utilizam estrutura de ramificação chamada nó, semelhante aos galhos ou à raiz de uma árvore, para representar uma decisão. São recomendados para mapear diversas possibilidades de resultados de uma decisão. Cada nó da árvore de decisão apresenta uma probabilidade de um resultado possível ocorrer.

As campanhas de marketing utilizam abusivamente esse algoritmo, em que uma forma comum é fornecer um desconto para a compra de um produto considerando a divisão dos clientes em três segmentos: um que provavelmente comprará não receberá desconto; outro que precisará de mais um estímulo para comprar recebe um pouco de desconto; e aquele com poucas chances de compra recebe um desconto maior. Pode, inclusive, segmentar um grupo que age negativamente a esse tipo de abordagem de venda e, claro, esse não receberá nenhuma mensagem. A partir dessa ramificação, o marketing pode organizar melhor a campanha e fazer aprimoramento com base nas reações do cliente.

**Redução de dimensionalidade** – Como o nome já identifica, reduz a dimensão dos dados removendo aqueles com pouca significância para análise. Esses algoritmos removem dados *outliers*, redundantes, sem significados e outros dados não úteis. Um exemplo bem comum é de um conjunto de dados proveniente de sensor de monitoramento de uma área. Nele estão contidos milhares de dados informando que um sensor está em operação. Para a análise, a informação de que o sensor está em operação não é relevante e pode atrapalhar o analista e a visualização do resultado, portanto, ela é identificada e retirada do conjunto de dados.

**Régressão linear** – Esse algoritmo é fundamental para o aprendizado de máquina e tem como base a análise estatística. Ajuda a modelar o relacionamento entre os dados por meio da quantificação da força de correlação entre variáveis do conjunto de dados. Pode também prever valores futuros.

Para melhor compreender, vamos supor que tenhamos uma tabela que apresenta ao longo de cinco anos a quantidade vendida de um modelo de carro juntamente com seu valor unitário. Se quisermos saber quantos carros serão vendidos para um determinado valor unitário, podemos utilizar a regressão linear para fazer uma previsão. A regressão linear analisa os dados e monta a equação de uma curva ou reta para, a partir daí, prever a localização dos próximos pontos considerando a correlação de valor e quantidade.

**Regularização para evitar *overfitting* (sobreajuste)** – O conceito de *overfitting* na estatística é utilizado para nomear um modelo que reflete perfeitamente os dados analisados, contudo, ele não é capaz de prever novos resultados. Esse algoritmo procura modificar o modelo para evitar *overfitting* (sobreajuste), tem como vantagem poder ser aplicado em qualquer situação de aprendizado de máquina. Essa técnica, geralmente, é utilizada para simplificar modelos estatísticos complexos e de previsões imprecisas, pois tais modelos são imprevisíveis quando exposto a um novo conjunto de dados diferente ao anterior. A técnica deve ser usada para avaliar a qualidade de um modelo, o analista de dados não deve se prender somente à medição do erro.

**Aprendizado de máquina baseado em regras** – Uma das principais características dos algoritmos de aprendizado com base em regras é que se utilizam de regras relacionais para descrever os dados, isso o diferencia dos algoritmos de aprendizado de máquina que criam um modelo que pode ser usado em várias situações. Esse algoritmo tem bons resultados com todos os dados que recebe. Contudo, ressalta-se que as dezenas de regras incluídas podem tornar o modelo muito complexo. À medida que o sistema é treinado (operacionalizado), exceções ao modelo vão surgindo. Nesses casos, é bom ter cuidado para que o modelo não fique complexo e impreciso. Consideramos como exemplo um modelo para o cálculo automático do imposto de renda para pessoa jurídica ou sistema tributário brasileiro.

**Aprendizagem de conjunto** – É uma técnica de implementação que considera o uso de um conjunto de algoritmos de aprendizado para obter um resultado preditivo melhor quando comparado com qualquer algoritmo de aprendizado isoladamente. Isso é possível pois a análise dos dados gera uma nova hipótese que geralmente não está presente nos modelos que a constituíram. Essa flexibilidade pode servir para ajustar os demais modelos.

## TEORIA EM PRÁTICA



Vamos ver o que é o algoritmo Convolutional Neural Networks, conhecido como ConvNets. Esses algoritmos levaram o resultado de análise com imagem a um nível muito alto. Eles trabalham com a verificação e identificação facial sem restrições. Essa técnica está sendo estudada extensivamente nos últimos anos, pois suas aplicações práticas são enormes. Os algoritmos mais recentes e de melhor desempenho de verificação de faces representam faces com recursos extremamente completos. Modelos de aprendizado profundo, como o ConvNets, são eficazes para extrair recursos visuais de alto nível e são muito usados nas aplicações de verificação de face. Os ConvNets são ensinados para classificar as faces disponíveis pelas suas identidades. Cada ConvNet recebe um caminho facial como entrada e extraí recursos profundos das camadas inferiores. No processo, os números de recurso vão reduzindo ao longo dos níveis mais baixos de extração de recursos, enquanto, ao mesmo tempo, os recursos mais globais e de alto nível são formados nos níveis superiores. No final do último nível, as informações da identidade são ricas e diretamente preditas com inúmeras classes de identidade, podendo chegar a milhares e, assim, identificar uma face.

Para mais detalhes, consulte Sun et al. (2014).

Diante da apresentação do ConvNet, pesquise outros exemplos de algoritmos e suas aplicações e tente criar possíveis usos.

Escolha um exemplo e estude como foi possível chegar a ele.

## VERIFICAÇÃO DE LEITURA

- 
1. Desenvolver um algoritmo ou escolher um já pronto é uma tarefa complexa, envolvendo, muitas vezes, análise estatística e matemática. Muitos algoritmos demonstram bom desempenho comprovado por análise técnica, porém outros funcionam bem considerando a experiência do contexto (SEDGEWICK, 1983). Ainda sobre algoritmos, é INCORRETO afirmar que um algoritmo é eficaz quando:
    - a. O procedimento reduz problemas a uma série de etapas de memorização a serem seguidas.
    - b. Fornece uma resposta certa e nunca uma resposta errada.
    - c. Possui um número finito de etapas.
    - d. Trabalha diversos tipos de problemas.
    - e. O procedimento reduz problemas com uma série infinita de etapas.
  2. De acordo com o conceito de *machine learning* (aprendizado de máquina), qual afirmativa abaixo é correta?

- I. Existem várias semelhanças entre o aprendizado de máquinas e de animais.
- II. A definição de *machine learning* está relacionada com aquisição de conhecimento, habilidade em algo, compreensão, instrução, estudo e experiência.
- III. Uma máquina aprende quando colhe amostras da sua fala e torna-se capaz de identificar você para uma interação.
- a. I, II e III.
- b. II e III.
- c. I e III.
- d. Somente a II.
- e. I e II.
3. A abordagem NÃO supervisionada é indicada quando o problema a ser analisado possui enorme quantidade de dados não estruturados ou sem rótulo.

POR TANTO

É utilizada em aplicativos de mídia social que utilizam a internet como meio de comunicação, exemplo: Facebook, Instagram, Pinterest, Twitter, Snapchat e outros.

Com base nas informações dadas e na relação proposta entre elas, é CORRETO afirmar que:

- a. As duas afirmações são verdadeiras e a segunda justifica a primeira.
- b. As duas afirmações são verdadeiras e a segunda não justifica a primeira.

- c. A primeira afirmação é verdadeira e a segunda é falsa.
- d. A primeira afirmação é verdadeira e a segunda é falsa.
- e. A primeira afirmação é falsa e a segunda é verdadeira.

## ► Referências bibliográficas

BROWNLEE, J. **A Tour of Machine Learning Algorithms**. Disponível em: <https://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/>. Acesso em: 4 abr. 2019.

METZ, J. MONARD, M. C. Clustering hierárquico: uma metodologia para auxiliar na interpretação dos clusters. In: XXV Congresso da Sociedade Brasileira de Computação. **Anais...** São Leopoldo-RS: Sociedade Brasileira de Computação, 2005.

NILSSON, J. N. **Introduction to machine learning**. Robotics Laboratory Department of Computer Science Stanford University Stanford, CA. 1998. Disponível em: <<https://ai.stanford.edu/~nilsson/MLBOOK.pdf>>. Acesso em: 30 ago. 2019.

O'CONNELL, A.; FRICK, W. **From data to Action**. HBR – Harvard Business Review, 2014. Disponível em:< [https://www.sas.com/content/dam/SAS/en\\_us/doc/whitepaper2/hbr-from-data-to-action-107218.pdf](https://www.sas.com/content/dam/SAS/en_us/doc/whitepaper2/hbr-from-data-to-action-107218.pdf)>. Acesso em: 30 ago. 2019.

SEGEWICK, R. **Algorithms**. Library of Congress Cataloging in Publication Data. Eua. 1983. Disponível em: <<http://dsp-book.narod.ru/Algorithms.pdf>>. Acesso em: 30 ago. 2019.

SMOLA, A.; VISHWANATHAN, S. V. N. **Introduction to Machine Learning**. Cambridge University Press, 2010. Disponível em: <<https://www.cse.iitb.ac.in/~ganesh/noml2015/bookchaptersvn.pdf>> . Acesso em: 30 ago. 2019.

SUN, Y.; WANG, X.; TANG, X. **Deep Learning Face Representation from Predicting 10,000 Classes**. IEEE Xplore. 2014. Disponível em: <[http://mmlab.ie.cuhk.edu.hk/pdf/YiSun\\_CVPR14.pdf](http://mmlab.ie.cuhk.edu.hk/pdf/YiSun_CVPR14.pdf)> . Acesso em: 30 ago. 2019.



## ► Gabarito

### Questão 1 – Resposta E

**Resolução:** a afirmação está INCORRETA, todo algoritmo tem o pressuposto de ser finito, ou seja, precisa finalizar o processamento e apresentar um resultado.

### Questão 2 – Resposta A

**Resolução:** as três afirmações estão corretas e fazem parte da definição e compreensão do conceito de machine learning.

### Questão 3 – Resposta A

**Resolução:** as duas afirmações são verdadeiras e se justificam, o ambiente de rede social é ideal para abordagens de aprendizado de máquina NÃO supervisionado.



# ***Cloud computing e Big Data***

Autor: Aimar Martins Lopes

## **► Objetivos**

- Compreender o que é *cloud computing*.
- Capacitar e analisar os modelos de serviços e implantação de nuvem.
- Descrever a relação da computação em nuvem com o Big Data.
- Conhecer os desafios do Big Data na computação em nuvem.

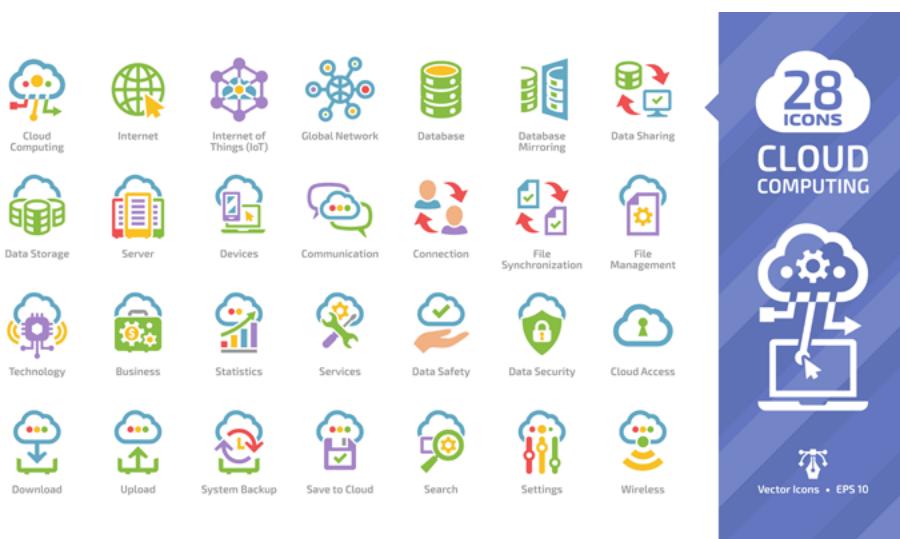
## ► 1. Computação em nuvem (cloud computing)

O termo computação em nuvem, em inglês *cloud computing*, ou simplesmente *cloud*, é utilizado por profissionais e usuários finais para tudo que pode ser conectado dinamicamente de forma onipresente e conveniente à *web service* (serviços de internet). O serviço compartilha um conjunto de recurso que pode ser utilizado e configurado com esforço mínimo (NIST, 2019). A computação em nuvem é caracterizada por acesso global, mobilidade, infraestrutura, plataforma padronizada, escalabilidade e gerenciamento de serviços.

Uma definição bem abrangente para computação em nuvem se refere ao modelo de negócio que tem como base a tecnologia da informação, que fornece serviço pela internet com o uso de hardware e software configurado sob demanda pelos clientes, independente do dispositivo de acesso, localização, escala dinâmica, qualidade, provisionamento rápido, compartilhamento, virtualização e interação (MADHAVI, TAMILKODI, JAYA, 2012).

A Figura 24 mostra alguns elementos fornecidos pela computação em nuvem.

Figura 24 – Computação em nuvem



Fonte: Yuriy Bucharskiy / iStock

Os diversos serviços prestados, geralmente, são pagos de acordo com seu uso, ou seja, *pay-per-use*. Como os recursos são flexíveis, o usuário pode reduzir ou aumentar o uso de forma rápida e fácil, pagando pelo uso. O fornecedor deve ter critérios claros de medição. O aumento ou a redução é por autosserviço, o usuário final é quem faz o acesso e configura o serviço.

Os recursos da nuvem, aos quais chamaremos de serviços, possuem as estruturas computacionais de hardware e software instalados em locais distantes dos usuários finais e operados por fornecedores especializados. O acesso é realizado, principalmente, por *web browser*, mas também por dispositivos móveis.

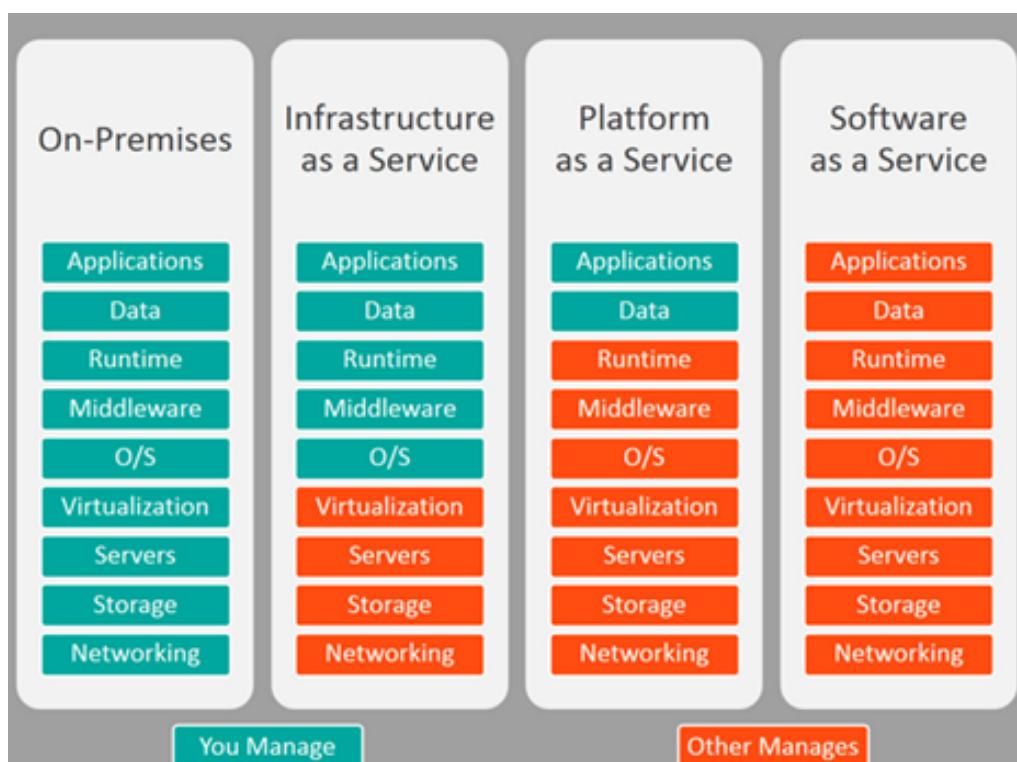
Os modelos de serviços são quatro:

- SaaS – Software as a Service, fornecimento de software (aplicações) e dados como serviço na internet: podem ser acessados de qualquer lugar e não há risco de perda de informação se seu computador quebrar ou acabar a energia.
- PaaS – Platform as a Service, fornecimento de plataforma como serviço: estruturas de desenvolvimento de aplicações são disponibilizadas para desenvolvedores construírem novas aplicações. Não há necessidade da compra de hardware e software, o ambiente disponibilizado possui as ferramentas necessárias para o ciclo de vida da aplicação, do desenvolvimento até a entrega.
- IaaS – Infrastructure as a Service: os fornecedores, chamados de provedores, disponibilizam recursos computacionais, tais como servidores, armazenamento, software, rede e espaços físicos, que são pagos de acordo com a demanda.

- DaaS – Desktop as a Service, *desktop* virtual como serviço: mais incomum que os demais, o usuário final pode hospedar serviços de computador individual na nuvem de computação. Por ser mais incomum, não trataremos dele aqui.

Esses modelos serão vistos com mais detalhes na Figura 25, que apresenta uma definição básica deles.

Figura 25 – Computação em nuvem – diferentes modelos de serviços



Fonte: <https://www.lastline.com/blog/how-cloud-computing-enables-and-threatens-organizations-digital-transformation/>.

## ASSIMILE

Para a AWS (Amazon Web Service), um dos maiores provedores de computação em nuvem do mundo, os fundamentos da computação em nuvem oferecem acesso rápido a recursos de TI flexíveis com baixo investimento e custo mensal, não há atividades de manutenção e

gerenciamento da infraestrutura. O cliente escolhe o serviço e tamanho correto de recursos computacionais necessários. Todos os recursos estão quase instantaneamente disponíveis, e só é pago o que foi usado (AWS, s/d).

A computação em nuvem pode ser implantada de formas variadas, privada, pública ou híbrida, conforme será abordado a seguir.

## 1.1 Modelos de serviços na nuvem

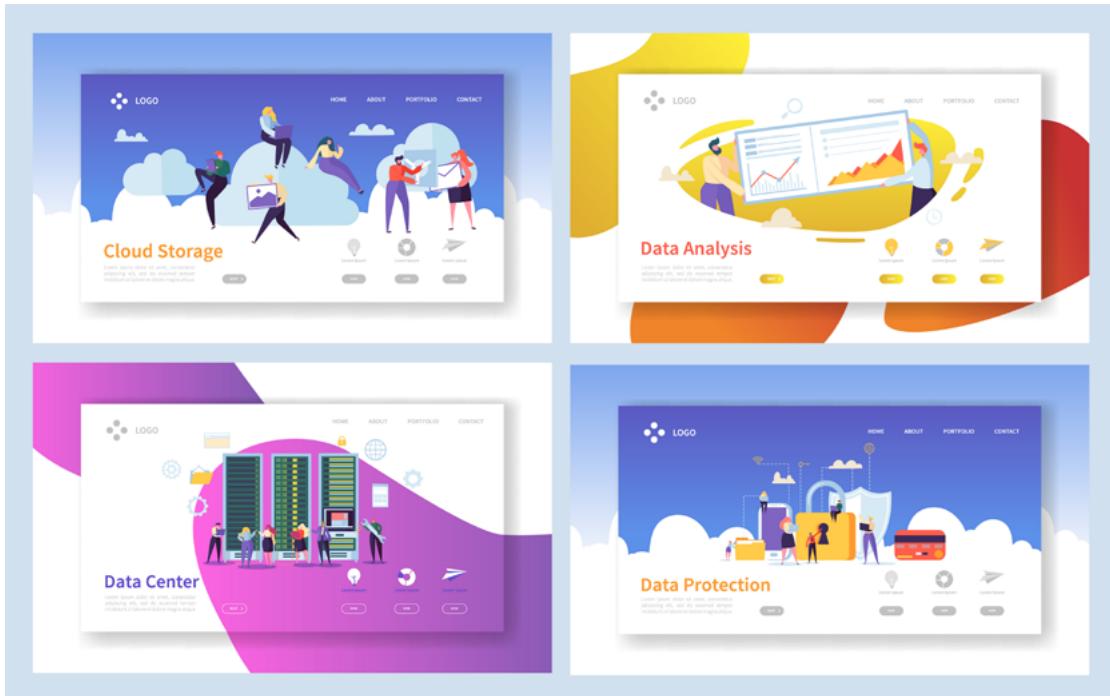
Por serem os mais comuns, utilizados e complexos, abordaremos com um pouco mais de detalhes o SaaS, PaaS e IaaS.

**SaaS** – Software como serviço, corresponde ao serviço mais próximo do usuário final. Os aplicativos são executados na nuvem, podendo ser configurados pelo cliente de acordo com sua necessidade. A esse processo de configuração, denomina-se customização. Os aplicativos e demais serviços devem estar disponíveis para acesso via navegador *web browser* 24h por dia. Os provedores que disponibilizam acesso aos serviços de vários tipos de dispositivos possuem mais vantagens competitivas.

À medida que o usuário necessita de mais ou menos serviço ou acrescentar um novo, isso pode ser realizado de forma transparente pelo provedor, como também toda a manutenção e atualização dos aplicativos tornam as atividades fáceis de executar. Além dessa escalabilidade, estão inclusos o arquivamento dos dados, a segurança e o fornecimento das licenças de software. A Figura 26 ilustra alguns serviços de computação de nuvem.

Exemplo: Salesforce, Dropbox, e-mails diversos, AWS Marketplace, ERP – ERPflex, Google Apps, Microsoft Sharepoint e muitos outros, inclusive os aplicativos legados das organizações que são disponibilizados na nuvem.

Figura 26 – Serviços de computação em nuvem



Fonte: invincible\_bulldog / iStock

**PaaS** – Plataforma como Serviço, geralmente é referenciada como um serviço de camada intermediária do modelo de computação em nuvem. É composta de hardware, computador virtual e software utilizado para o desenvolvimento de aplicações, entre eles banco de dados, sistema operacional, armazenamento de dados, serviços de comunicação, *framework* de desenvolvimento, etc.

Esse serviço facilitou e tornou acessível ferramentas a custo competitivo para os desenvolvedores. Não é necessário adquirir hardware nem software para desenvolvimento, teste, implantação, compartilhamento de banco de dados, hospedagem e manutenção de aplicativos. A grande vantagem competitiva está na integração e no compartilhamento sob medida oferecidos pela internet.

Exemplo: Windows Azure, Google App Engine, IBM Softlayer, Oracle Cloud.

**IaaS** – Infraestrutura como serviço é a base do sistema computacional, composta por hardware e software que sustentam as plataformas e os aplicativos, baseada na virtualização dos recursos computacionais escalados dinamicamente. Toda essa infra pode estar espalhada em várias regiões geográficas. Alguns provedores resumem esse serviço em fornecimento de armazenamento e processamento. O usuário não precisa se preocupar com manutenção, compra, depreciação, segurança, espaço físico, etc.

Como os demais serviços descritos, o IaaS possui escalabilidade dos serviços com o pagamento do que foi utilizado, dispensa a despesa e complexidade da aquisição e o gerenciamento de servidores e infraestrutura de *data center*.

Exemplo: Amazon EC2, Rackspace hosting, EMC, Eucalyptus, IBM cloud.

O Amazon Elastic Compute Cloud (Amazon EC2) é um serviço da AWS de processamento computacional sob demanda, projetado para computação em escala na nuvem. A interface de configuração é simples e de fácil operação. O serviço Amazon EC2 permite escalabilidade, em minutos pode-se inicializar novas instâncias de servidor, tanto para mais ou para menos, de acordo com a necessidade. O cliente paga somente o serviço utilizado. No pacote de serviço do Amazon EC2 também estão incluídas ferramentas para criar aplicativos resistentes a falhas (AMAZON, 2019).

## 1.2 Modelos de implantação na nuvem

A computação em nuvem também é classificada de acordo com sua implantação. Os modelos mais utilizados são nuvem privada, pública e híbrida. Vamos conhecer um pouco melhor cada um deles.

**Nuvem Privada** – Neste modelo, a infraestrutura pode estar instalada remotamente ou local, pode ser alugada ou proprietária e pode ser gerenciada por terceiros ou pela organização usuária, contudo, toda a infraestrutura é dedicada para uma única organização. Quando a operação e o controle de acesso são da equipe de TI da organização, a qualidade do serviço e a segurança também é de sua responsabilidade.

Este modelo encarece e dificulta a operação contínua dos serviços, no entanto, a organização detém todo o controle da configuração da nuvem para atender a seus requisitos específicos. São indicados por terem maior flexibilidade e segurança. Os usuários mais comuns são bancos, grandes multinacionais, governo, organizações de operação crítica, etc.

**Nuvem pública** – Neste modelo, a infraestrutura pertence ao provedor dos serviços para o público em geral. São ofertados por rede pública de comunicação de dados, geralmente internet, e podem ser acessados por quem desejar utilizar. Os serviços são caracterizados por componentes de TI instalados, gerenciados e disponibilizados pelo provedor, possuem configurações padrão genéricas, em sua maioria carecendo de restrições reguladas e segurança elevada.

Os recursos são compartilhados pelos usuários de mesmo serviço, não são exclusivos como na nuvem privada. Os serviços podem ser acessados por todos os funcionários da organização que os contratou de qualquer lugar. Exemplo: MS Azure, serviços de e-mail, Google, AWS – Amazon, etc.

**Nuvem híbrida** – A nuvem híbrida usa a nuvem privada em combinação com a pública, possibilitando um ambiente de vantagem de ambas as opções que pode ser aproveitado pela organização. Quando a organização usa a nuvem privada, ocorre um isolamento dos demais recursos de TI, o que vai demandar cargas de trabalho entre base de dados, aplicativos, etc.

A estratégia adotada pelas organizações é disponibilizar na nuvem privada as aplicações críticas, as de dados mais sensíveis, as que requerem mais segurança, as financeiras e, na nuvem pública, as de baixa segurança, como consulta de site, e-mail, etc. Complementam a estratégia, o custo, a segurança, disponibilidade, velocidade e os recursos disponíveis.

Resumidamente, o uso da computação em nuvem se justifica, principalmente, por redução de custo, aumento de produtividade e conveniência. Todos os dados e aplicativos ficam armazenados na nuvem e não no notebook ou desktop, ou ainda no celular. Pela internet, o usuário encontra a conveniência de acessar e utilizar os serviços com diferentes aplicativos, em diferentes locais, sem se preocupar com *backups* ou *atualização de aplicativo*, tudo isso quem faz é o fornecedor da nuvem.

## ► 2. A relação da computação em nuvem com o Big Data

A computação em nuvem permite acesso de recursos onipresente, de acordo com a necessidade e demanda de recursos de computação que podem ser configurados rapidamente com pouco esforço de gestão do provedor de serviços.

O Big Data, para cumprir seus objetivos, requer uma gama de recursos poderosos, servidores, banco de dados, ferramentas que coletam, classificam e processam grande volume e variedade de dados em formatos diferentes com alta velocidade. O Big Data é categorizado por classes que reúnem seus componentes. A compreensão dessa estrutura facilita o entendimento da relação da computação em nuvem com Big Data. No Quadro 9 abaixo, apresentamos as cinco classes do Big Data.

## Quadro 9 – As cinco classes do Big Data



Fonte: elaborado pelo autor.

O Big Data rastreia tarefas de forma paralela pela internet e aloca os resultados em sistema de armazenamento, por exemplo, o HDFS (sistema de arquivos distribuídos, exemplo Hadoop), isso é adequado aos modelos de serviços SaaS, IaaS e PaaS das nuvens. Após essa etapa, os dados são mapeados, ou seja, são divididos em processos paralelos distribuídos, com o objetivo de simplificar os problemas. Na última etapa, os resultados das análises são reduzidos e reunidos novamente no HDFS. Um *framework* muito conhecido que realiza essa tarefa é o MapReduce (DHABHAI; GUPTA, 2016).

Resumidamente, o uso da computação em nuvem no Big Data pode ser descrito como tendo início nas fontes de dados provenientes

de bancos de dados e web armazenados em sistemas distribuídos tolerante a falhas (Hadoop), processamento paralelo em larga escala distribuído (MapReduce), análise dos dados e relatórios e, finalmente, a visualização por meio de diferentes gráficos e formas para a tomada de decisão. São exemplo de Big Data em plataforma de nuvem: Google Cloud Services, MS Azure e AWS S3. Conclui-se que as tecnologias da computação em nuvem, Big Data e internet estão intimamente ligadas.

Os serviços de Big Data têm crescido rapidamente na computação em nuvem, principalmente pelo uso da tecnologia de virtualização. Toda a tecnologia envolvida na computação em nuvem é adequada para a oferta de um modelo de serviço. Ela é eficiente e eficaz e, definitivamente, é um padrão para integração e fornecimento das inúmeras tecnologias utilizadas para análise de dados.

### ► 3. Alguns desafios do Big Data

Com o crescimento rápido de novas tecnologias e o já estabelecido serviço de computação em nuvem e Big Data, é primordial que as organizações implementem e amadureçam continuamente o uso dessas duas tecnologias. Os executivos do alto escalão hierárquico das organizações têm sofrido pressão para adoção dessas tecnologias, pois, além de uma questão estratégica de aproximação dos clientes e criação de modelos de negócios associado ao Big Data, há o custo alto de manter as estruturas computacionais legadas, custo esse que a computação em nuvem pode reduzir.

Contudo, o Big Data também precisa romper barreiras; desafios relacionados à demanda cada vez maior ou armazenamento e processamento em alta escala implicam riscos e consequências

graves. A seguir, apresentamos alguns pontos relacionados com tais desafios.

**Disponibilidade dos recursos na nuvem** – A solicitação, modificação e atualização dos recursos da nuvem são realizadas de forma rápida, ou seja, os recursos sob demanda estão à disposição. O desafio está no crescimento da demanda de armazenamento e processamento relacionado com a capacidade do provedor e da indústria da tecnologia da informação conseguir atender à demanda no futuro.

**Escalabilidade do armazenamento** – O sistema de armazenamento de dados distribuído e escalável na nuvem é um ponto crítico. A escalabilidade se refere à capacidade de armazenar o volume crescente de dados de maneira confiável. Para isso, as tecnologias de banco de dados são: NoSQL para dados não estruturados (chave-valor, em memória, documento, gráfico e pesquisa) e SQL ou DBMS (Database Management System) para dados estruturados.

**Integridade de dados** – Mantém-se a preocupação que há com os bancos de dados legados que estão na infraestrutura da organização, o que muda é que, no conceito de computação em nuvem, eles estão em algum lugar remoto. A integridade dos dados se refere a perda, modificação, acesso e exatidão. A estrutura da nuvem deve garantir a segurança dos dados, gerenciar os usuários, fornecer proteção física e controlar o acesso.

**Privacidade** – Associado a integridade dos dados na questão do armazenamento, a segurança e a privacidade na computação em nuvem também é um desafio. As informações sobre os níveis de acordo de serviço equivocados (SLA – Service Level Agreement), mais ataques aos sistemas por hackers de diferentes partes do mundo, ataques esses que podem violar dados ou retirar os sistemas do ar, são preocupantes. Para tanto, os provedores de serviços na nuvem instalam sistemas com tolerância a falhas.

**Qualidade dos dados coletados** – O Big Data colhe dados de várias fontes, o problema é que muitas não são conhecidas nem confiáveis e não há maneiras de verificação na maioria delas. Portanto, o volume e a velocidade de crescimento das fontes de dados passam a ser outro desafio, dado ruim vai gerar informação ruim.

**Variedade dos dados** – Apesar de ser uma das características do Big Data, a variedade dos dados e as suas diversas fontes de origem caracterizam um desafio. Com a expansão, a dificuldade de armazenar e tratar dados não estruturados e semiestruturados aumenta, pois eles são indicados para estrutura de armazenamento simples, com pouca hierarquia e muito flexibilidade.

**Preparação de dados** – Após coletar e reunir os dados, é preciso fazer a preparação para que estejam no mesmo formato estrutural. Como as fontes estão cada vez mais diversas, essa preparação está cada vez mais difícil. Seguem algumas fontes: dispositivos móveis, site, blog, rede social, sensores, texto, IoT, imagens e vídeos. Portanto, a preparação, purificação, análise e transposição desses tipos de dados para posterior armazenamento são desafiadoras.

**Análise de Big Data** – A escolha do profissional, bem como os modelos adequados para análise de grandes dados, é perigosa. São necessários profissionais capacitados, comprometidos e de confiança para elaborar os modelos de análise, bem como ferramentas modernas. Como se não bastasse, a taxa de fluxo de dados é constante e ininterrupta em alguns modelos. Nesses casos, é preciso distinguir e montar intervalos para o armazenamento.

**Segurança de Big Data** – As ameaças à segurança são constantes. Os serviços de computação em nuvem precisam de criptografia, algoritmos para gerenciar as chaves de segurança e mecanismo para trocar essas chaves entre os sistemas das organizações envolvidas.



## PARA SABER MAIS

O Apache Hadoop é um *framework* (estrutura) que executa aplicativos em grandes conjuntos de hardware (clusters). A estrutura é composta de aplicativos que trabalham de forma transparente e confiável o processamento de dados. Possui uma técnica computacional chamada Mapear/Reducir (MapReduce). No processamento, o aplicativo é dividido em pequenos fragmentos de código e executado na nuvem em qualquer parte do cluster de hardware. Os dados são armazenados em um sistema de arquivos distribuídos (HDFS), tanto o armazenamento quanto o processamento são realizados de forma distribuída e paralela. O *framework* também é projetado para ser tolerante a falha, ou seja, a falha é corrigida automaticamente (APACHE HADOOP, s/d).



## TEORIA EM PRÁTICA

Vamos ver um caso de mineração de dados por meio de um serviço de computação em nuvem.

Esta mineração foi realizada em uma base de dados do Twitter na nuvem com o uso dos serviços da AWS. Computação em nuvem usada para analisar grandes quantidades de dados no Twitter. O algoritmo Page Rank foi utilizado para obter rankings de usuários da base do Twitter. Os serviços de infraestrutura, armazenamento e processamento de nuvem da AWS foram usados. A hospedagem de todos os cálculos relacionados também. O processo computacional foi realizado em duas fases: na primeira, fase de rastreamento, todos os dados foram recuperados das bases do Twitter; na segunda, o

processamento feito pelo algoritmo Page Rank analisava os dados coletados. Durante a fase de rastreamento, foi observado a geração de 50 milhões de nós e 1,8 bilhão de bordas, que significava aproximadamente dois terços da base de usuários do Twitter. Diante do volume, esse é um exemplo de solução relativamente barata para aquisição e análise de dados com o uso de uma infraestrutura de nuvem (HASHEM et al., 2014).

Page Rank é um algoritmo para classificar e posicionar sites de acordo com a busca na internet. Qualifica de acordo com a quantidade e qualidade da busca.

Realize uma leitura sobre o algoritmo Page Rank para melhor compreensão de seu funcionamento.

## VERIFICAÇÃO DE LEITURA

1. O termo computação em nuvem, em inglês *cloud computing*, ou simplesmente *cloud*, é comum ser utilizado por profissionais e usuário final a tudo que pode ser conectado dinamicamente à *web service* (serviços de internet). Sobre a computação em nuvem, é INCORRETO afirmar:
  - a. A computação em nuvem é caracterizada por acesso global, mobilidade, infraestrutura, plataforma padronizada, escalabilidade e gerenciamento de serviços.
  - b. Um modelo de negócio baseado em tecnologia da informação, que provê serviço pela internet com o uso de hardware e software utilizado sob demanda.

- c. Os diversos serviços prestados, geralmente, são pagos de acordo com seu uso, ou seja, pay-per-use.
  - d. Os recursos são fixos, o usuário não pode reduzir ou aumentar o uso de forma rápida e fácil.
  - e. O aumento ou a redução é por autosserviço, o usuário final é quem faz o acesso e configura o serviço.
2. De acordo com os modelos de serviços na nuvem, qual afirmativa abaixo é correta?
- I. SaaS – Software como serviço: corresponde ao serviço mais próximo do usuário final. Os aplicativos são executados na nuvem, podendo ser configurados pelo cliente de acordo com sua necessidade.
  - II. PaaS – Plataforma como serviço: é uma camada intemediária do modelo de computação em nuvem, composta de hardware, computador virtual e software utilizado para o desenvolvimento de aplicações.
  - III. IaaS – Infraestrutura como serviço: é a base do sistema computacional, composta por hardware e software que sustentam as plataformas e os aplicativos, baseada na virtualização dos recursos computacionais escalados dinamicamente.
- a. I, II e III.
  - b. II e III.
  - c. I e III.
  - d. Somente a II.
  - e. I e II.
3. A computação em nuvem permite acesso de recursos onipresente, de acordo com a necessidade e demanda

de recursos de computação que podem ser configurados rapidamente.

## POR TANTO

O Big Data requer uma gama de recursos poderosos, servidores, banco de dados, ferramentas que coletam, classificam e processam grande volume e variedade de dados em formatos diferentes com alta velocidade. Por isso, todo seu processamento deve ser local, sem o uso de serviços de nuvem.

Com base nas informações dadas e na relação proposta entre elas, é CORRETO afirmar que:

- a. As duas afirmações são verdadeiras e a segunda justifica a primeira.
- b. As duas afirmações são verdadeiras e a segunda não justifica a primeira.
- c. As duas afirmações são falsas.
- d. A primeira afirmação é verdadeira e a segunda é falsa.
- e. A primeira afirmação é falsa e a segunda é verdadeira.

## ► Referências bibliográficas

AMAZON EC2. Disponível em: <https://aws.amazon.com/pt/ec2/>. Acesso em: 24 abr. 2019.

APACHE HADOOP. Disponível em: [https://wiki.apache.org/hadoop#Apache\\_Hadoop](https://wiki.apache.org/hadoop#Apache_Hadoop). Acesso em: 23 abr. 2019.

AWS. **O que é a computação em nuvem.** Disponível em: <https://aws.amazon.com/pt/what-is-cloud-computing/>. Acesso em: 22 abr. 2019.

DHABHAI, A.; GUPTA, Y. K. A Study of Big Data in Cloud Environment with their Related Challenges. **IJESC**, Índia, v. 6, n. 8., 2016.

HASHEM, I. A. T.; YAQOOB, I.; ANUAR, N.B.; MOKHTAR, S.; GANI, A.; KHAN, S.U. The rise of “Big Data” on cloud computing: review and open research issues. **Information Systems**, v. 47, p. 98-115, 2015.

MADHAVI, K. V.; TAMILKODI, R.; JAYA, S. K. Cloud Computing: Security threats and Counter Measures. **IJRCCCT**, v. 1, n. 4, p. 125-128, 2012.

NIST. **Nist publica versão final de definição de *cloud computing*.** Disponível em: <https://chapters.cloudsecurityalliance.org/brazil/2011/11/18/nist-publica-versao-final-de-definicao-de-cloud-computing/>. Acesso em: 7 mai. 2019.

## ► **Gabarito**

### **Questão 1 – Resposta D**

**Resolução:** a afirmação está INCORRETA, pois os recursos são flexíveis, o usuário pode reduzir ou aumentar o uso de forma rápida e fácil, pagando pelo uso.

### **Questão 2 – Resposta A**

**Resolução:** as três afirmações estão corretas e compõem os modelos de serviço da computação em nuvem.

### **Questão 3 – Resposta D**

**Resolução:** a primeira afirmação é verdadeira, pois a computação em nuvem permite o acesso de qualquer lugar, a qualquer tempo. A segunda afirmação é falsa, pois afirma que o Big Data não pode ser executado na nuvem. No entanto, essa é a plataforma mais utilizada para ele. As demais afirmações do Big Data estão corretas.



# Estruturas de programação em nuvem

Autor: Aimar Martins Lopes

## ► Objetivos

- Compreender o que é computação em nuvem (*cloud computing*).
- Capacitar e analisar os modelos de serviços em nuvem.
- Conhecer exemplo de serviço em nuvem.
- Caracterizar e diferenciar linguagem de programação.

## 1. Computação em nuvem

A computação em nuvem, oferecida no modelo IaaS, é caracterizada por acesso global, mobilidade, infraestrutura, plataforma padronizada, escalabilidade e gerenciamento de serviços disponíveis na internet, ou seja, a infraestrutura de que a organização necessita, bem como seu gerenciamento, também realizado por terceiros. A organização não tem a preocupação com servidor, processamento, armazenamento de dados, escalabilidade ou qualquer outra infraestrutura.

Essa infraestrutura é fornecida pelos serviços de SaaS – Software as a Service, PaaS – Platform as a Service e IaaS – Infrastructure as a Service. A partir dessa condição, o desenvolvimento de aplicativos (software) passa a ser realizado na nuvem com o uso de PaaS, o desenvolvedor não necessita de grande poder computacional, pois toda a necessidade de edição, programação, compilação, execução e disponibilidade de uso é provida pelo serviço na nuvem.

A ideia central da programação em nuvem é o fornecimento de software pela internet (SaaS), com o uso de plataforma (PaaS) para desenvolvimento dos softwares e gerenciamento dos serviços, acesso e armazenamento dos dados (IaaS), lembrando que os serviços são executados por terceiros e o mercado oferece várias formas de atendimento (JULA, 2014).

Exemplo de PaaS para desenvolvimento de aplicativos são Google Cloud Platform, Windows Azure, IBM Blue Cloud, Joyent (Samsung), Salesforce, AWS Cloud9, Aptana Cloud, Oracle Cloud e outros. Geralmente esses ambientes suportam as mesmas linguagens de desenvolvimento, Java, JavaScript, Python, PHP, .NET, HTML5, Node.js, Spring, Ruby, etc.



## ASSIMILE

a plataforma como serviço (PaaS) é constituída por computador, equipamentos de hardware, linguagens, sistemas operacionais, ferramentas de desenvolvimento de aplicativos e interfaces de usuário. O provedor fornece requisitos básicos configurados, incluindo o funcionamento do sistema, rede, servidores para o programador desenvolver e disponibilizar aplicativos na nuvem (JULA et al., 2014).

Qual seria a vantagem de adotar a programação em nuvem? Além de todos os benefícios da infraestrutura, a computação em nuvem permite elasticidade de forma rápida e segura, pode-se expandir ou ativar novos serviços rapidamente. Por exemplo, em semana de promoção da Black Friday, os consumidores acessam os sites de *e-commerce* muito mais vezes. Para manter a qualidade do serviço e não perder vendas, os fornecedores rapidamente configuram a infraestrutura de atendimento.

Para o desenvolvimento na nuvem, usa-se *framework*. Nele, geralmente está presente a tecnologia IDE (Integrated Development Environments), que pode ser entendida como um conjunto de aplicativo (software) disposto em uma plataforma de ambiente integrado de desenvolvimento, reunindo recursos para programação, tais como: editor de linguagem, compilador, interpretador ou depurador, etc. A diferença e vantagem é que essa tecnologia permite independência de dispositivo de acesso específico, ou seja, a programação pode ser feita de qualquer lugar e dispositivo. A Figura 27 apresenta uma ilustração dos recursos da computação em nuvem.

Figura 27 – Recursos da computação em nuvem



Fonte: LisLud / iStock

## ► 2. Arquitetura da computação em nuvem

A evolução da computação em nuvem provém da evolução de várias tecnologias, desde as físicas relacionadas com hardware até as lógicas relacionadas com os algoritmos e softwares.

Na década de 1980, a ascensão dos microcomputadores e sua evolução na década seguinte também impulsionou a modernização dos *mainframes*. Anos seguintes, a área de rede local e global se desenvolveram juntamente com a tecnologia cliente-servidor, inclusive com servidor hospedado na internet. Nos anos 2000, a internet se expandiu por todo o mundo e em seguida incorporou várias tecnologias, como o HTTP (Hypertext Transfer Protocol), SMTP (Simple Mail Transfer Protocol), o sistema em hipermídia WWW (World Wide Web), linguagem (HTML, Java, PHP, Python, ASP e outras), armazenamento (MySQL, Postgres SQL, Oracle, Microsoft SQL Server), *frameworks*

de desenvolvimento, API (Application Programming Interface), processamento distribuído e outras tecnologias que permitiram replicar o ambiente local em um ambiente remoto. Essas tecnologias integradas permitem soluções mais produtivas, seguras com compartilhamento de recursos e menor custo.

Outras tecnologias compõem esse grupo e caracterizam o ambiente da computação em nuvem. A seguir, apresentamos algumas delas:

**Arquitetura SOA (Service-Oriented Architecture)** – Nessa arquitetura, as aplicações fornecem suas funcionalidades como serviço na web, por meio de processamento distribuído/paralelo, instruções de requisição e replicação (*request/reply*) entre as aplicações e as conexões dos serviços da web, chamadas *web services*. Nessa arquitetura, os desenvolvedores, para construir o ambiente, geralmente utilizam o protocolo SOAP (Simple Object Access Protocol), a linguagem XML (Extensible Markup Language) e outros.

Outra tecnologia é formada pelas linguagens de programação – várias linguagens com diversas características estão à disposição. As aplicações podem ser desenvolvidas com *frameworks* que facilitam a programação, teste e disponibilidade.

**Virtualização** – Técnica utilizada para separar um ambiente de aplicação e sistema operacional do ambiente físico (hardware). Pode ser aplicada em máquina local ou na nuvem. Uma máquina virtual é caracterizada pela escalabilidade e flexibilidade, o cliente configura e paga sob demanda.

**Fluxo de serviço** – É o conjunto completo das ofertas sob demanda disponíveis na computação em nuvem ofertada com base em diferentes tecnologias de forma automática e transparente. Tem como principal objetivo fornecer um fluxo de trabalho contínuo para o cliente.

**Armazenamento distribuído** – Sem dúvida, a tecnologia responsável pela consolidação de outras, pois com o armazenamento e a recuperação de dados em servidores espalhados pela rede, torna-se possível que os clientes acessem suas informações e gerenciem os serviços sem esforço e com transparência.

### PARA SABER MAIS



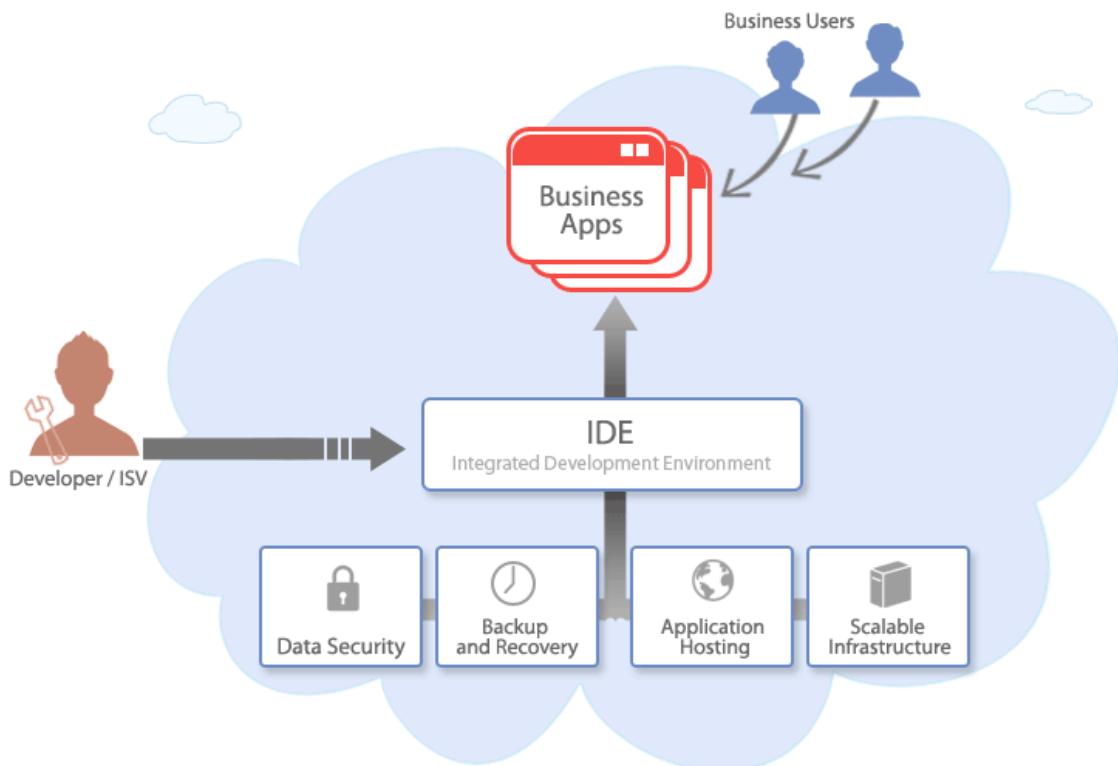
A característica elasticidade, primordial na computação em nuvem, também é conhecida como escalabilidade, capacidade de escalar ou reduzir serviços e recursos na nuvem. O cliente altera a configuração do serviço sempre que necessita. Essa flexibilidade torna os serviços atraentes, pois, com boa gestão, a redução de custo é inevitável. Essa característica é adotada pelos provedores mais competitivos da computação em nuvem, o da Amazon chama-se EC2 (Elastic Compute Cloud).

## ► 3. Ambiente do desenvolvedor

O ambiente de desenvolvimento para o programador está oferecido no serviço PaaS. Os provedores criam ambiente de desenvolvimento de aplicações que atende às necessidades dentro da internet (nuvem). Esse ambiente é composto de ferramentas de *layout* de aplicação, linguagens, interpretadores e outros suficientes para que uma aplicação seja desenvolvida e colocada em uso pelo serviço SaaS. A esse ambiente ou plataforma, costuma-se denominar *framework*. A vantagem é que o programador não precisa de computadores poderosos nem banco de dados local para o desenvolvimento, assim ele consegue aumentar a produtividade e mobilidade, ou seja, pode produzir em ambientes

mais criativos e confortáveis. A Figura 28 ilustra um ambiente de desenvolvimento, comumente chamado de *framework*.

Figura 28 – Fluxo de programação na internet



Fonte: ZOHO, 2019.

O desenvolvimento pode ser realizado em ambiente de software livre (*open source*), geralmente mais simples, sem a necessidade de grande habilidade de programação por parte do desenvolvedor, pois não são muito sistematizados; o Eclipse é um exemplo. Contudo, também há ambiente integrado bem sofisticado com recursos integrados (IDE) em uma só aplicação, ou seja, editor de texto, linguagem, recurso de teste (depuração do programa), gerenciamento de versões e projeto, por exemplo, o MS – Visual Studio.

A Eclipse Foundation define sua atuação como a plataforma aberta de inovação e colaboração de provimento global para sua comunidade e organizações por meio de um ambiente maduro, escalável e

comercialmente amigável de software aberto de colaboração e inovação (ECLIPSE, s/d).

A Microsoft anuncia em sua página que possui o melhor conjunto de ferramentas para desenvolvedor, formado por: IDE Visual Studio, Code Visual Studio, Azure DevOps e Visual Studio App Center (MICROSOFT, s/d.(b))

## ASSIMILE

Um *framework* reúne estruturas semelhantes de linguagens de programação diferentes em um único software, com o intuito de criar uma função genérica. Seu principal objetivo é facilitar e aumentar a produtividade do desenvolvedor em um projeto. Contudo, é necessário ter um bom domínio da linguagem nativa do *framework*, a qual está programando, pois cada *framework* é baseado em uma linguagem de programação, exemplo: Java (Hibernate e Spring), PHP (Zend e Laravel), CSS (Bootstrap), Python (Django), Ruby (Ruby on Rails), Java Script (Angular JS) e C# (Asp .net).

## ► 4. Microsoft AZURE

É uma plataforma de computação em nuvem para prover serviços especializados para desenvolvedor em ambiente Microsoft. Oferece recursos para desenvolvimento de software como também ambiente de execução tanto na nuvem como fora dela. Os recursos disponibilizados são:

- Sistema operacional Windows Azure é plataforma de nuvem para desenvolvimento e execução de serviços, sistemas e aplicações. Oferece o serviço baseado na computação em cluster e virtualização.
- .NET é a plataforma de desenvolvimento e execução de aplicativos e sistemas.
- SQL Services é um sistema de banco de dados executado em uma plataforma em nuvem com acesso fornecido por meio de serviço.
- Live Services são os serviços oferecidos, seguem alguns: processamento, máquina virtual com suas funcionalidades, ferramentas de rede, armazenamento e bancos de dados, aplicações web móvel, container, ferramenta de análise de dados, serviços de Inteligência Artificial e IoT.
- *Sharepoint* é a plataforma de aplicações que uma organização utiliza para intranet, internet, site, gestão e disponibilização de portal e conteúdo, disponibilizar sistemas, etc.
- Dynamics é o sistema de gestão empresarial ERP (Enterprise Resource Planning), integra as funções organizacionais e auxilia na tomada de decisão do executivo.

Na estrutura, os desenvolvedores podem criar aplicativos, compilar e personalizar aplicativos disponíveis na nuvem. Da mesma maneira que uma macro é criada no Excel, a plataforma fornece ferramentas com componentes de software para programação. O ambiente é mais produtivo e reduz o tempo de codificação, pois contém componentes pré-codificados, tais como: fluxo de trabalho, serviços de diretório, recursos de segurança (MICROSOFT, s/d.(a)).

Segundo o site da Microsoft, o Microsoft Azure traz diversos serviços de nuvem que dá liberdade para o cliente desenvolver, gerir e disponibilizar

aplicativos na internet, com estrutura e ferramenta que mais deseja. Na lista de serviços encontra-se DevOps, IoT, análise de dados, *framework* e outros, dispostos na modalidade de demanda necessária com custo competitivo. (MICROSOFT, s/d.(a)).

Quadro 10 – Arquitetura do Windows Azure



Fonte: elaborado pelo autor, adaptado de Microsoft (s/d.(a)).

## ► 5. Google Cloud Platform – GCP

A Google nasceu com a criação de produtos para a internet atuando globalmente. Dentre o enorme portfólio de soluções, encontramos o Google Cloud, que é composto de vários serviços, sendo o GCP (Google Cloud Platform) e o G Suite os principais. O primeiro é um conjunto de programas com processamento em nuvem que oferece várias plataformas computacionais, dentre elas servidores, discos rígidos para armazenamento, máquinas virtuais, banco de dados, ambiente de desenvolvimento, etc. O segundo oferece versões personalizadas de produtos independentemente do domínio do cliente, tais como: agenda, e-mail, drive virtual, Hangouts, site, planilha, documento, etc.

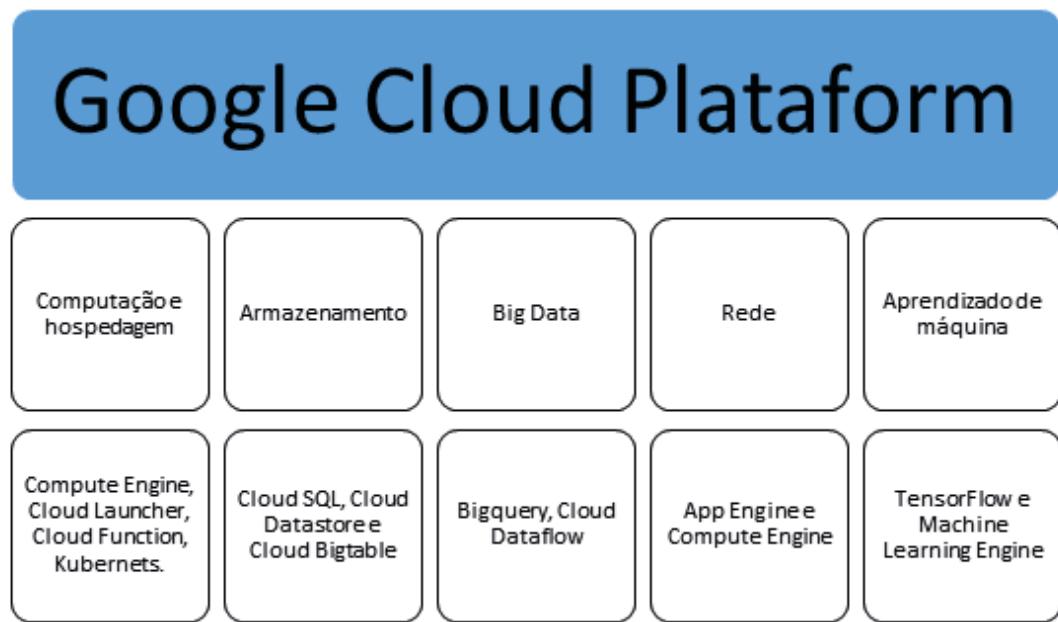
Os serviços seguem as regras da computação em nuvem, onde o hardware e software são fornecidos por meio de serviços sob demanda. Os serviços acessam diversos recursos necessários para a entrega final, também podem ser agrupados até a composição do serviço final. Os serviços do GCP mais usados são: computação e hospedagem, armazenamento, rede, Big Data e aprendizado de máquina (GOOGLE, s/d.(b)).

- O serviço de computação e hospedagem fornece processamento, gestão de aplicativos, tecnologias de container e possibilita que o cliente crie sua própria infraestrutura. Veja alguns:
  - Execução de funções sem servidor no Google Cloud Functions, no serviço de plataforma (PaaS) Google App Engine e gerencia os recursos feitos pelo cliente;
  - O SDK do App Engine permite o desenvolvimento e teste de máquina;
  - O Google Cloud SQL fornece vários bancos de dados de terceiros;
  - APIs e bibliotecas do cliente são gerenciadas no Cloud Endpoints;
  - Segurança fica por conta do Cloud Security Scanner;
  - O Google Kubernetes Engine faz a computação no ambiente container;
  - Máquina virtual do Google Compute Engine e o Google Cloud Launcher.
- O serviço de armazenamento Cloud SQL é oferecido em banco de dados relacional MySQL, PostgreSQL ou não relacional NoSQL em

Cloud Datastore e Cloud Bigtable. O armazenamento pode ser em redundância geográfica, em alta disponibilidade, com baixo ou alto custo, dependendo da estratégia. O cliente também pode optar por instalar o banco de dados de sua preferência.

- O serviço App Engine de gerenciamento de rede configura redes, firewall e trajetos, cria registros DNS, gerencia o tráfego entre os recursos contratados e conecta redes do cliente à rede do Google. O Compute Engine conecta diversos recursos para a solução de um projeto, o firewall faz o controle do tráfego. Com esses recursos, o cliente pode criar VPN (Virtual Private Network) e trajetos específicos para a aplicação.
- Os serviços de Big Data fornecem processamento e consulta aos dados na nuvem. A análise de dados é de responsabilidade do BigQuery, que transforma os dados de origens diversas em tabelas, tem comandos SQL, interface gráfica, comando de API e outros. O processamento de dados em lote ou *streaming* é feito pelo Cloud Dataflow, destinado para grande volume de dados e processamento paralelo. Realiza tarefas de ETL (extração, transformação e carga), migração e transformação de dados em outro sistema de armazenamento.
- Os serviços de *machine learning* e inteligência artificial são avançados, contém APIs pré-treinadas que podem ser utilizadas pelos clientes ou aplicativos específicos disponíveis. O serviço é gerenciado pelo TensorFlow. O Cloud Machine Learning Engine é o serviço para treinar os modelos de aprendizado de máquina e as APIs prontas para uso são: API Google Cloud Video Intelligence, API Google Cloud Speech, API Google Cloud Vision, API Google Natural Language, API Google Cloud Translation, Dialogflow Enterprise Edition (GOOGLE, s/d.(b))

Quadro 11 – Arquitetura do GCP



Fonte: elaborado pelo autor.

## ► 6. Linguagem de programação

A linguagem de programação é um recurso básico dos sistemas e aplicativos, é claro que esses não existiriam na sua ausência. Uma linguagem é utilizada pelo desenvolvedor (programador de computador) para escrever os códigos do software (programa). O código é um conjunto de instruções simbolizadas e estruturadas que, posteriormente, é transformado em uma linguagem que o computador entende, chamamos essa linguagem de linguagem de máquina.

Devido à importância das linguagens na computação em nuvem, falaremos de algumas delas.

**JavaScript** - O JavaScript é uma linguagem de alto nível interpretada que todo desenvolvedor precisa conhecer, está presente em quase todo software desenvolvido para a nuvem. Sua principal característica é ser interpretada, leve no processamento, compatível com a maioria

dos navegadores de internet e flexível. Sua utilização está centrada no desenvolvimento de aplicações próximas ao usuário final.

**Python** - O Python é uma linguagem de programação de alto nível, de fácil aprendizado e uso. Caracterizada por sintaxe intuitiva e limpa. Permite uma programação clara e seu uso está nas mais diversas áreas, desde aprendizado de máquina, análise de dados até aplicativo comercial. Devido a sua versatilidade, é a linguagem que mais tem crescido no mundo.

**Java** - O Java, linguagem de alto nível orientada a objeto, revolucionou o uso da internet, hoje de propriedade da Oracle, mas criada por cientistas da Sun Microsystems. É boa para desenvolvedor iniciante pela sua flexibilidade e seu grande uso na área comercial. É comum encontrarmos em sistemas de indústria, comércio, banco, etc. Seu uso é tanto para aplicações próximas do usuário (*front end*) como para aplicações de retaguarda (*back end*) que residem no servidor. É indicada para desenvolvimento de grandes sistemas pela sua confiabilidade e estabilidade.

A linguagem é conhecida no meio dos desenvolvedores da tecnologia da informação como a linguagem de um único código, justificada pela sua adaptabilidade de execução em vários ambientes. Isso é possível pois seus objetos de execução fazem referência somente a dados internos e alocação automática de memória. Java é compatível com a maioria dos *frameworks*.

**C e seus sabores** - A linguagem C tornou-se popular por volta de 1970, devido a sua robustez e confiabilidade, ganhou uma série de variedades, C++, C# entre outras. É uma linguagem que contribui muito para a programação, quando a necessidade é alto desempenho, com certeza ela será recomendada. O sistema operacional Linux tem sua base de código em C e C++, esta última orientada a objeto, que por sua vez também foi codificada em C.

Apesar de ser de alto nível, portanto clara e estruturada, sua maior força é o desempenho. Não é à toa que os desenvolvedores a utilizam para

desenvolver jogos, realidade virtual, computação gráfica e outras que necessitam de alta performance.

**PHP** – A sigla da linguagem PHP é tão utilizada que seu significado quase nunca é mencionado. Originalmente conhecida por Personal Home Page e agora por Hypertext Preprocessor, é indicada para a criação de página dinâmica e interativa na web. Está presente na maioria dos sites e em sites de grandes organizações. O fato de não ser tão estimada, talvez seja justificado pelo seu objetivo, ser uma linguagem de propósito geral, executada por qualquer servidor, programação fácil com script para página HTTP, fácil de usar e configurar, além de ser gratuita. Contudo, carrega a desvantagem de ser lenta e reduzir o desempenho do site.

**TypeScript** – Essa linguagem de código aberto foi criada em 2012 pela Microsoft. Sua principal característica está no conjunto rigoroso de sintaxe do JavaScript, portanto, é muito semelhante a ela, o que a torna conhecida e de fácil uso. Seu uso tem crescido rapidamente e é indicada para aplicativos de larga escala, desenvolve aplicativos tanto do lado do usuário como do servidor.

**Ruby** – Linguagem de programação de alto nível criada para simplificar e aumentar a produtividade do ambiente de desenvolvimento, a fim de torná-lo leve. É dinâmica e seu código é aberto. Está presente em muito dos aplicativos, sendo a base do *framework* Ruby on Rails. Como é simples e sem regras rígidas, o desenvolvedor consegue criar um aplicativo com pouca linha de código, mas tem como ponto negativo a falta de flexibilidade.

**SQL** – A linguagem SQL (Structured Query Language) foi criada para manipular operações em bancos de dados, opera com as funções de recuperação, manipulação e armazenamento de dados em um banco de dados relacional. É muito utilizada para extrair e explorar dados no auxílio da tomada de decisão pela alta gerência empresarial.



## TEORIA EM PRÁTICA

A estratégia da ContaAzul era não ter estrutura interna de hardware e software para suportar e operacionalizar os serviços do ERP. O desafio era prover um ERP num serviço de qualidade com uma equipe reduzida, disponibilidade de 24h e autonomia na configuração da infraestrutura e boa relação custo-benefício. O provedor de computação em nuvem escolhido para enfrentar o desafio, foi a AWS – Amazon Web Service, pois ofertou escalabilidade para o crescimento da ContaAzul com baixo impacto no valor dos serviços, além de uma plataforma que suporta as linguagens Java, Python, Ruby e Shell. Atualmente, para desenvolvimento e provimento dos serviços, a ContaAzul tem contratado um conjunto de serviços da AWS, tais como: EC2 (Amazon Elastic Compute Cloud), EBS (Amazon Elastic Block Store), RDS (Amazon Relational Database Service), SQS (Amazon Simple Queue Service), Amazon CloudWatch, S3 (Amazon Simple Storage Service), Amazon Glacier, AWS Elastic Beanstalk, Amazon DynamoDB e Amazon ElastiCache (AMAZON, s/d.).

Sugiro realizar pesquisa na internet e leitura de outros casos para a melhor compreensão do potencial da computação em nuvem.



## VERIFICAÇÃO DE LEITURA

1. A tecnologia IDE (Integrated Development Environments) pode ser entendida como um conjunto de aplicativo (software) extremamente poderoso para desenvolvedores. Sobre a tecnologia IDE nuvem, é INCORRETO afirmar:

- a. É um conjunto de aplicativo (software) disposto em uma plataforma de ambiente integrado de desenvolvimento.
  - b. Reúne recursos para programação, tais como: editor de linguagem, compilador, interpretador ou depurador.
  - c. A tecnologia permite independência de dispositivo de acesso à aplicação.
  - d. A programação de aplicativo pode ser feita de qualquer lugar e dispositivo.
  - e. A tecnologia é poderosa, mas limita o local de programação e acesso pelo usuário.
2. De acordo com a arquitetura dos serviços na nuvem, qual afirmativa abaixo é correta?
- I. Na arquitetura SOA (Service-Oriented Architecture), as aplicações fornecem suas funcionalidades como serviço na web, por meio de processamento distribuído/paralelo, instruções de requisição e replicação (*request/reply*).
  - II. Na PaaS, várias linguagens com diversas características estão à disposição. As aplicações podem ser desenvolvidas com *frameworks* que facilitam a programação, o teste e a disponibilidade.
  - III. Virtualização é caracterizada como a técnica de máquinas virtuais com escalabilidade e flexibilidade, em que o cliente configura e paga sob demanda.
- a. I, II e III.
  - b. II e III.
  - c. I e III.
  - d. Somente a II.
  - e. I e II.

3. O Google Cloud é formado por um conjunto de programas com processamento em nuvem que oferece várias plataformas computacionais, dentre elas, servidores, discos rígidos para armazenamento, máquinas virtuais, banco de dados, ambiente de desenvolvimento.

PORTANTO

Quando o cliente contrata o Google Cloud, ele tem acesso ao ambiente interno de sua organização para configurar os recursos de acordo com a necessidade e demanda de computação.

Com base nas informações dadas e na relação proposta entre elas, é CORRETO afirmar que:

- a. As duas afirmações são verdadeiras e a segunda justifica a primeira.
- b. As duas afirmações são verdadeiras e a segunda não justifica a primeira.
- c. As duas afirmações são falsas.
- d. A primeira afirmação é verdadeira e a segunda é falsa.
- e. A primeira afirmação é falsa e a segunda é verdadeira.

## ► Referências bibliográficas

AMAZON. **Estudo de Caso ContaAzul.** Disponível em: <https://aws.amazon.com/pt/solutions/case-studies/conta-azul/>. Acesso em: 1 mai. 2019.

AMAZON EC2. Disponível em: <https://aws.amazon.com/pt/ec2/>. Acesso em: 24 abr. 2019.

ECLIPSE. **The Platform for Open Innovation and Collaboration.** Disponível em: <https://eclipse.org/>. Acesso em: 29 abr. 2019.

GOOGLE. **Google Cloud Platform.** s/d.(a). Disponível em: <https://cloud.google.com/products/?hl=pt-br>. Acesso em: 30 abr. 2019.

\_\_\_\_\_. **Sobre os serviços do GCP.** s/d.(b). Disponível em: <https://cloud.google.com/docs/overview/cloud-platform-services?hl=pt-br>. Acesso em: 30 abr. 2019.

JULA, A.; SUNDARARAJAN, E.; OTHMAN, Z. Cloud computing service composition: A systematic literature review. **Expert Systems with Applications**, v. 41, p. 3.809-3.824, 2014.

MICROSOFT. **Microsoft Azure.** s/d.(a). Disponível em: <https://azure.microsoft.com/pt-br/overview/what-is-azure/>. Acesso em: 29 abr. 2019.

\_\_\_\_\_. **Visual Studio.** s/d.(b). Disponível em: <https://visualstudio.microsoft.com/pt-br/>. Acesso em: 29 abr. 2019.

ZOHO. Disponível em: <https://www.zoho.com/creator/images/subpages/paas.gif>. Acesso em: 21 mai. 2019.

## **Gabarito**

### **Questão 1 – Resposta E**

**Resolução:** a afirmação está INCORRETA, pois, ao contrário da afirmação, a IDE é flexível quanto ao local de programação e acesso pelo usuário.

### **Questão 2 – Resposta A**

**Resolução:** as três afirmações estão corretas no conceito dos serviços.

### **Questão 3 – Resposta C**

**Resolução:** a segunda afirmação é falsa, pois o Google Cloud é um serviço de computação em nuvem de forma onipresente, os recursos estão disponíveis para acesso e configuração remoto e não local.



**Bons estudos!**