

INTRODUÇÃO A BIG DATA E INTERNET DAS COISAS (IOT)



Izabelly Soares de Moraes

Ciência de dados e *Big Data*

Objetivos de aprendizagem

Ao final deste texto, você deve apresentar os seguintes aprendizados:

- Descrever dados e *datasets*.
- Discutir ciência de dados e *Big Data*.
- Listar práticas de ciência de dados e *Big Data*.

Introdução

As informações e os dados nunca foram tão acessíveis quanto o são hoje em dia. Por meio da internet, conseguimos saber basicamente de tudo que ocorre na nossa localidade e no mundo. A cada ação nossa, provavelmente, muitos dados estão sendo gerados para as empresas responsáveis pelos artefatos e serviços tecnológicos de que fazemos uso.

Neste capítulo, você vai compreender melhor sobre os conceitos de dados e *datasets*, assim como vai conseguir visualizar a ação conjunta que pode haver entre a ciência de dados e as tecnologias *Big Data*, tanto por meio de contextualizações quanto de práticas.

Dados e *datasets*

Você já parou para pensar na quantidade de observações que estamos sempre fazendo em tudo que está ao nosso redor? O ser humano, devido à sua racionalidade, consegue lidar com, interpretar e associar diversos acontecimentos quase que simultaneamente. E essa não é uma característica desenvolvida apenas quando somos adultos. Se você entrar em uma loja de brinquedos com uma criança, você compreenderá melhor o que estamos falando, já que as primeiras reações quase sempre serão as de as escolhas serem realizadas com base em alguns padrões já preestabelecidos pela criança. Mas você pode estar se questionando sobre o que isso tem a ver com dados e *datasets*, não é?

Basicamente tudo, pois, quando fazemos observações sobre algo, no decorrer do tempo, vamos formando padrões, até mesmo definindo preferências, e

agimos dessa forma durante toda a nossa vida. Além disso, geralmente, nossas escolhas são baseadas nessas experiências. Se fizermos uma analogia com essa situação comum do cotidiano com o mundo dos negócios, em que decisões são tomadas a todo instante, não seria muito diferente, tendo em vista que todo negócio constrói um conhecimento sobre si mesmo e sobre seus clientes e produtos no decorrer do tempo.

Hoje, ao acompanhamos pesquisas e noticiários, deparamo-nos com um protagonista que já existe há muito tempo, mas que só dos últimos tempos para cá virou o foco de todos: os dados. Mas como podemos defini-los?

A definição mais básica de um dado é sabermos que, se estiver só, ele não faz sentido, de modo deve haver informações sobre ele, ou seja, complementos informacionais e até mesmo contextos, para que ele tenha sentido e possa gerar algum conhecimento.

Ao analisarmos a fundamentação do conceito de dados, vemos claramente que ele é um ativo importante dentro de um negócio, e podemos afirmar que nas nossas atividades cotidianas também! Você conseguiu perceber como somos geradores de dados e informações constantes?

Mas e *dataset*, o que seria? Em sua tradução livre, o termo significa conjunto de dados. Mencionamos que as informações são um coleção de dados e, dentro desse contexto, é relevante notar que o contexto científico exige que visualizemos níveis mais profundos dos processos dedutivos e intuitivos de observação para que possamos registrá-los com precisão. Uma maneira de fazer isso é construir um conjunto de dados, os quais são apresentados de várias formas. Em sua grande maioria, os dados são representados por meio de planilhas, podendo conter diversas linhas ou colunas, e não necessariamente precisam assumir aquela ideia que temos de planilhas desenvolvidas em alguns *softwares* específicos.

Um conjunto de dados possui algumas características relevantes, como a estruturação dos dados, já que, como citado anteriormente, lidar com dados é um trabalho extremamente minucioso, tendo em vista que o dado é o recurso chave de todo processo. Deve haver, também, a possibilidade de recuperação, acesso e identificação dos dados diante de todo o conjunto, ação que geralmente ocorre por meio de comandos ou disponibilização de links de acesso, além de certa frequência nas atualizações dos dados.

Do ponto de vista de Ramakrishnan e Gehrke (2013, p. 784), existem muitos motivos para que os dados sejam semiestruturados. A estrutura dos dados pode ser implícita, oculta, desconhecida ou o usuário pode optar por ignorá-la. Além disso, ao se integrar dados de várias fontes heterogêneas, a troca e a transformação de dados são problemas importantes. Dessa forma, é necessário que haja um

modelo de dados altamente flexível para integrar dados de todos os tipos de fontes, incluindo arquivos simples e sistemas legados.



Fique atento

Juntamente aos conceitos de dados, é importante compreendermos, também, os conceitos de informação e conhecimento.

- Informação: fornece significado para o dado, pois pode ser definida como um dado contextualizado.
- Conhecimento: pode receber diversas definições diferentes, mas, dentro desse contexto, o conhecimento pode ser definido como uma experiência ou até mesmo aprendizado obtido devido à organização e à contextualização dos dados e das informações.

O uso da ciência de dados e *Big Data*

Vivemos no século XXI e, no contexto dos negócios, as previsões acabam sendo o ponto forte das empresas. Mas não estamos falando de previsões de sorte, com cartomantes ou videntes; falamos de tecnologias que usam seus poderosos algoritmos para fazer previsões de negócios.

Quando falamos sobre essas estimativas, estamos ressaltando ainda mais a importância do dado, já que todo seu ciclo de vida (Figura 1) complementa o investimento em tecnologias tanto inseridas em ferramentas quanto em metodologias no mundo corporativo.

A concepção de um dado pode ser oriunda das mais diversificadas fontes: no nosso caso, produzimos dados em praticamente toda ação que desempenhamos por meio de algum recurso tecnológico, pois, quando nos conectamos à rede, nossas informações começam a ser obtidas imediatamente, seja por um *login* em algum *site* ou até mesmo pelas permissões que damos ao fazermos *download* e instalarmos algum aplicativo. Porém, algumas fontes de dados não podem estar susceptíveis a variações ou sofrer outros danos, ou seja, deve haver certa estrutura para receber os dados. Não podemos esquecer que existem regras empresariais que acompanham (ou pelo menos tentam acompanhar) todo esse processo. Por isso, por exemplo, existem tipos de dados que ficam armazenados por muito mais tempo que outros, e essa decisão cabe à empresa. Perceba que, quando falamos em armazenamento, estamos falando também de investimento em segurança computacional, e até mesmo em *hardware* e espaço na nuvem (*cloud computing*) capaz de armazenar tantas informações.

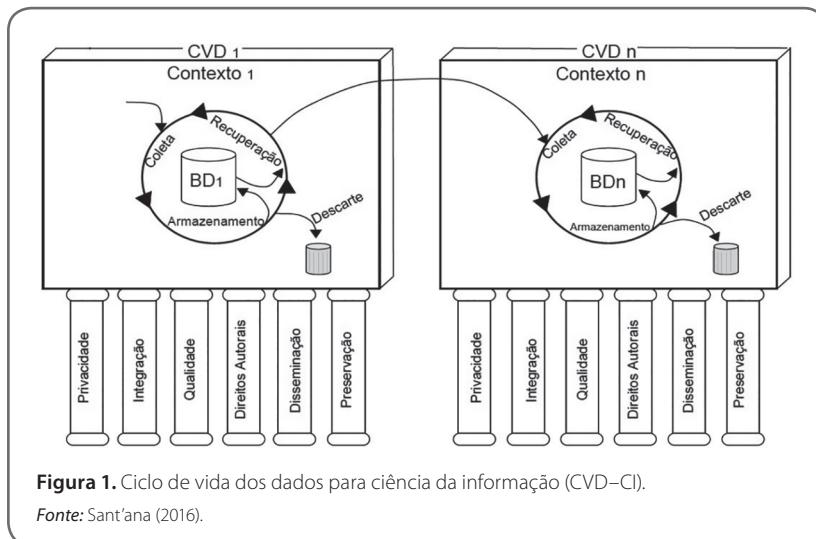


Figura 1. Ciclo de vida dos dados para ciência da informação (CVD-Cl).

Fonte: Sant'ana (2016).

Dessa forma, podemos concluir que a concepção, o armazenamento, o tratamento e outros processos relacionados aos dados irão variar conforme sua finalidade. Inclusive, pode haver até mesmo o descarte dos dados, como mostra a Figura 1.

A importância no conhecimento destes processos se dá, segundo Turban e Volonino (2013, p. 67), devido a três princípios:

- **1. Princípio da redução do valor de dados.** Uma análise dos dados em seu ciclo de vida mantém a atenção em como o valor dos dados pode diminuir de acordo com o seu envelhecimento. Assim, o dado tem mais valor quanto mais recente for. A maioria das organizações não pode operar no máximo de seu desempenho tendo pontos cegos, isto é, falta de dados disponíveis, de 30 dias ou mais.
- **2. Princípio do uso de dados 90/90.** Uma atuação em tempo real ou quase em tempo real sobre dados operacionais pode trazer vantagens importantes. De acordo com esse princípio, a maioria dos dados armazenados raramente é acessada após 90 dias (exceto para fins de auditoria). Ou seja, os dados perdem grande parte de seu valor após três meses.
- **3. Princípio de dados em contexto.** Para capturar, processar, formatar e distribuir dados rapidamente e quase em tempo real, é necessário um grande investimento em infraestrutura de gerenciamento de dados para fazer a ligação remota dos sistemas presentes nos pontos de venda

(PDVs) para armazenamento de dados, sistemas de análise de dados e aplicativos que geram relatórios. Esse investimento se justifica de acordo com o princípio de que dados devem estar integrados, ser processados, analisados e formatados em “informação acessível”. Os usuários finais precisam visualizar os dados em um formato significativo e em contextos, já que eles irão guiar suas decisões e seus planejamentos.

A ciência de dados transforma os dados utilizando conceitos matemáticos e estatísticos por meio de processos de mineração e filtragem dos dados. As ferramentas computacionais se tornam necessárias para, em conjunto com os *softwares*, realizar o armazenamento, a obtenção e o tratamento dos dados.

Mas e como tudo isso acontece? Para isso, são utilizadas tecnologias, como mencionamos anteriormente, e uma delas que podemos destacar aqui é *Big Data*. Como o próprio termo já sugere, isso significa lidar com uma grande quantidade de dados diversos (estruturados ou não estruturados).



Fique atento

Os dados não estruturados são aqueles dados cujo contexto total nem sempre a tecnologia consegue visualizar, como, por exemplo, em arquivos textuais.

Já os dados estruturados conseguem ser totalmente classificados e identificados com o uso das tecnologias.

O que não podemos deixar de comentar é que existem vários profissionais que lidam com os dados e que, muitas vezes, com a ajuda da tecnologia, nem sempre estão totalmente ligados ao setor de tecnologia da empresa, já que, na maioria das vezes, as tomadas de decisões são realizadas por profissionais administrativos.

Aplicações práticas da ciência de dados e *Big Data*

Várias metodologias estão sendo utilizadas pelas empresas para a coleta de dados. Porém, quando falamos em *Big Data*, estamos assumindo que, além de termos grande volume de dados, devido às grandes proporções, eles não podem ser tratados com métodos tradicionais. Para isso, devem ser executados alguns passos, tais como obtenção, armazenamento, sistematização e análise dos dados. O termo *Big Data* é, muitas vezes, caracterizado por três vertentes que,

inclusive, são conhecidas como os três Vs: **volume, variedade e veracidade**. Porém, é comum nos depararmos com outros 2 Vs que também contemplam de forma coerente os conceitos sobre essa tecnologia: o **valor** e a **velocidade**.

A obtenção dos dados pode ocorrer oriundas de diversas fontes, tanto internas quanto externas aos negócios da empresa. Já o armazenamento, provavelmente, ocorrerá por meio de sistemas e servidores. Isso ocorre para garantir que seja feito o armazenamento automático e para que possam ser realizados *backups* posteriormente.

As próximas etapas consistem na organização e na análise dos dados, nas quais deve haver um agrupamento que tenha como base algum padrão dos dados, gerando, com isso, uma estrutura capaz de facilitar o acesso e a análise das informações, que é basicamente o último passo a ser executado. Com a organização, isso fica mais fácil, já que é na extração que podemos obter a visualização dos dados úteis para as tomadas de decisão.

Ainda sobre a etapa de análise, podemos afirmar que, antes, acontecia apenas de forma descritiva, com o objetivo de trazer por meio, muitas vezes, de gráficos, planilhas e relatórios, alguns conjuntos de dados que caracterizavam as ações já executadas pela empresa. Porém, com a ciência de dados, essa etapa evoluiu, não só devido ao uso das tecnologias, mas também em relação a seus objetivos. Hoje, por exemplo, as análises não ocorrem apenas com olhares para o passado, mas também com perspectivas futuras, que são chamadas de análises preditivas e diagnósticas, já que é por meio desse processo que falhas ou pontos de melhoramentos são observados.

Existem diversas maneiras de as empresas coletarem dados, tais como: endereço de e-mail e IP, informações dos dispositivos, *browsers*, cliques em anúncios, seja pela rede social ou por e-mails, histórico de buscas, dentre outros. Para isso, elas utilizam alguns métodos para monitorar seus usuários, como ferramentas para identificação de dispositivos, perfis dos usuários, *cookies*, dentre outros.

Os dados obtidos das mais variadas fontes podem ser utilizados de diversas formas. As grandes empresas, por exemplo, utilizam para as integrações de contas, em que todas suas informações, muitas vezes, podem estar associadas ao seu e-mail, a conteúdos personalizados, e as empresas podem aumentar o leque de opções de produtos e serviços conforme cada perfil de cliente.

A associação desses recursos pode ser visualizada e utilizada em diversos contextos, como no esporte, em que diversos times de basquete, futebol, entre outros, passaram a utilizar a análise de dados para prever possíveis melhorias de seus times, tanto em resultados gerais quanto de desempenho dos atletas durante a prática de suas atividades.

Aqui no Brasil, até os órgãos governamentais utilizam *data warehouse* para registros. Nesse sentido, um setor atuante é o Ministério da Justiça, com o intuito de identificar ações de lavagem de dinheiro, dentre outros golpes. Empresas como Google, Facebook e as de tecnologia também utilizam fortemente a análise de dados, e podemos até afirmar que eles atuam por meio de todas as formas possíveis para isso, já que seus lucros são oriundos desse tipo de ação.

Grandes redes varejistas do setor alimentício também utilizam a ciência de dados e tecnologias *Big Data* para gerir seus negócios e, como mencionamos anteriormente, para ampliar a gama de negócios, produtos e serviços que podem ser disponibilizados aos clientes. Conforme Taurion (2013), *Big Data* é um conjunto de tecnologias, processos e práticas que permitem às empresas analisarem dados a que antes não tinham acesso e tomar decisões ou mesmo gerenciar atividades de forma muito mais eficiente. Para o autor, diversos dados podem ressaltar o início da curva de aprendizado sobre o que é o *Big Data* e seu respectivo impacto social. Como exemplo, ele cita o uso de *Big Data* pelas empresas Amazon e Netflix, que utilizam sofisticados e avançados sistemas de recomendação.

Portanto, as aplicações práticas que possuem ciência de dados e *Big Data* trazem maior transparência, já que os dados ficam disponíveis em locais específicos, e alguns deles até passam a ser de domínio público, como é o caso de dados governamentais. Ocorre, também, a amplificação das informações, tendo em vista que a conexão de diversas informações sobre algo específico ocorre de forma mais fácil.



Saiba mais

A análise de dados pode ser realizada com o uso de algumas tecnologias, como NoSQL, Hadoop, Sisence, TIBCO Spotfire, dentre outras.



Referências

RAMAKRISHNAN, R.; GEHRKE, J. *Sistemas de gerenciamento de banco de dados*. 3. Ed.

Porto Alegre: Penso, 2013.

SANTANA, R. C. G. Ciclo de vida dos dados: uma perspectiva a partir da ciência da informação.

Informação & Informação, v. 21, n. 2, p. 116–142, 2016. Disponível em: <<http://www.uel.br/revistas/uel/index.php/informacao/article/view/27940/20124>>. Acesso em: 23 dez. 2018.

TAURION, C. *Big Data*. Rio de Janeiro: Brasport, 2013.

TURBAN, E.; VOLONINO, L. *Tecnologia da informação para gestão: em busca do melhor desempenho estratégico e operacional*. 8. ed. Porto Alegre: Bookman, 2013.

Leituras recomendadas

AMARAL, F. *Introdução à ciência de dados: mineração de dados e big data*. Rio de Janeiro: Alta Books, 2016.

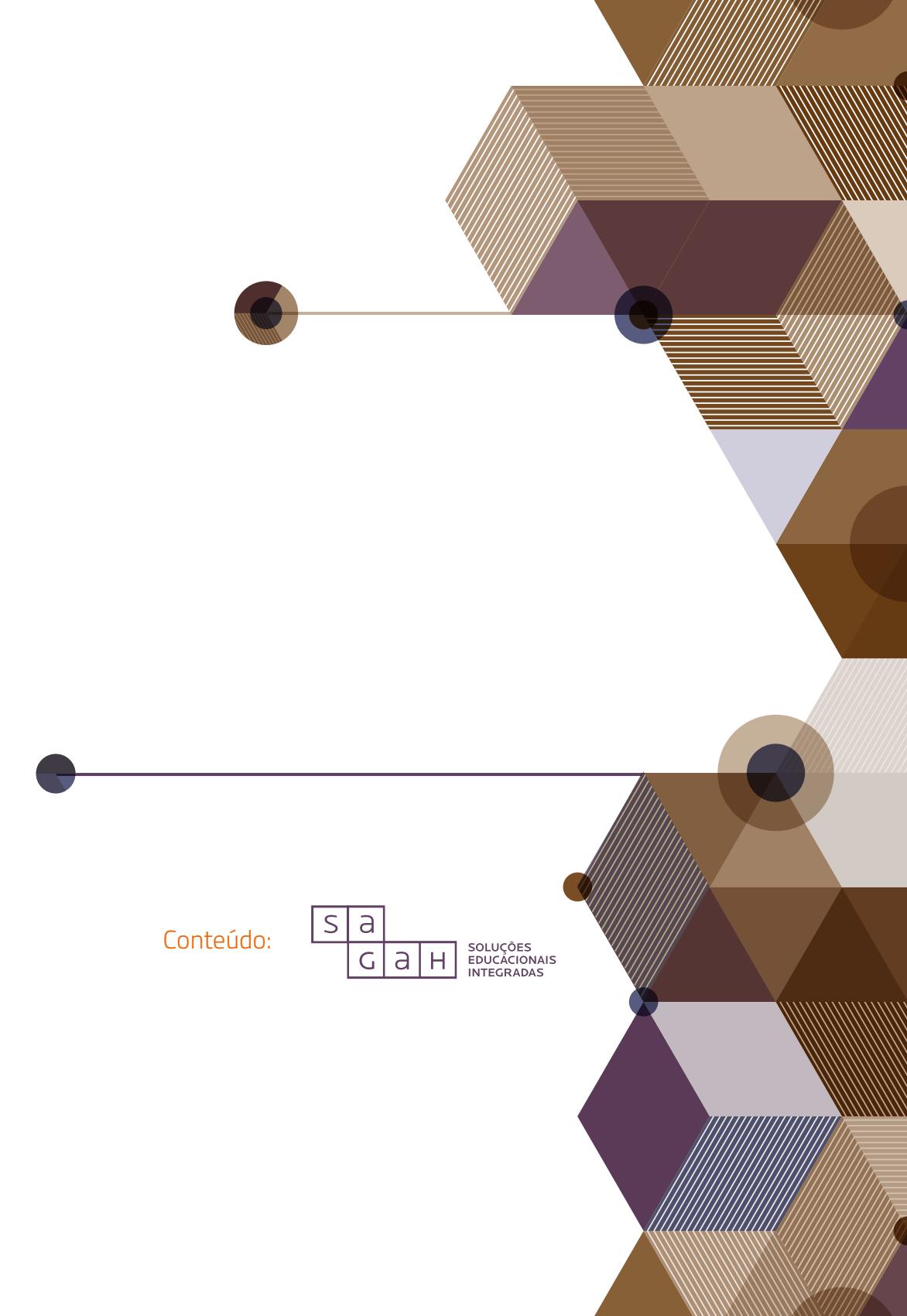
ANTONELLI, R. A.; NEITZKE, A. C. A.; VOESE, S. B. Relação entre a qualidade da informação recebida e o nível de tomada de decisão dos profissionais da área de negócios. *Revista Ambiente Contábil*, v. 10, n. 2, jul./dez. 2018. Disponível em: <<https://periodicos.ufrn.br/ambiente/article/view/12739/9538>>. Acesso em: 23 dez. 2018.

CAVIQUE, L. Big Data e Data science. *Boletim APDIO*, v. 51, p. 11-14, dez. 2014. Disponível em: <https://repositorioaberto.uab.pt/bitstream/10400.2/3918/1/2%20Boletim_51.11-14.pdf>. Acesso em: 23 dez. 2018.

FOREMAN, J. W. *Data Smart: usando data science para transformar informação em insight*. Rio de Janeiro: Alta Books, 2018.

PROVOST, F.; FAWCETT, T. *Data Science para negócios: o que você precisa saber sobre mineração de dados e pensamento analítico de dados*. Rio de Janeiro: Alta Books, 2016.

Encerra aqui o trecho do livro disponibilizado para esta Unidade de Aprendizagem. Na Biblioteca Virtual da Instituição, você encontra a obra na íntegra.



Conteúdo:



SOLUÇÕES
EDUCACIONAIS
INTEGRADAS