Big Data

De acordo com Salinas e Lemus (2017), o termo *Big Data* foi criado em 1997 por Michael Cox e David Ellsworth, pesquisadores da NASA que tinham que trabalhar com conjuntos de dados geralmente muito grandes, o que sobrecarregava a memória principal, disco local e capacidade de disco remoto. Eles chamaram isso de problema do *Big Data.*

Apesar de ser amplamente referenciado, *Big Data* não tem uma definição rigorosa e consensual. Geralmente está associado ao tratamento de dados massivos, extraídos de diferentes fontes e sem estruturas pré-definidas (SALINAS e LEMUS, 2017). De acordo com Gandomi e Haider (2015), cerca de 95% dos dados tratados por tecnologias de *Big Data* são dados não estruturados.

Para alguns autores, *Big Data* nada mais é do que um conjunto de dados cujo tamanho está além das ferramentas típicas de bancos de dados para capturar, armazenar, gerenciar e analisar (SALINAS e LEMUS, 2017). De acordo com SAS (2022), *big data* refere-se a conjuntos de dados tão grandes, rápidos ou complexos que são difíceis ou impossíveis de processar usando métodos tradicionais. O ato de acessar e armazenar grandes quantidades de informações para análise existe há muito tempo. Mas o conceito de *big data* ganhou força no início dos anos 2000.

Para Anand (2019), *Big data* é uma tecnologia utilizada para armazenar dados, tanto em formatos não estruturados quanto semi estruturados e estruturados, utilizando dispositivos de armazenamento mais baratos. Para agilizar o processamento, este é feito de forma descentralizada e distribuída por múltiplos servidores. Os dados são armazenados em formato nativo, sem um esquema ou modelagem definida.

Segundo Oussous et al. (2018) o termo *big data* refere-se a grandes conjuntos de dados, em constante crescimento, que incluem formatos heterogêneos de dados estruturados, não estruturados e semiestruturados. *Big data* possui natureza complexa e exige tecnologias sofisticadas e algoritmos avançados. Neste novo contexto, ferramentas tradicionais de *Business Intelligence* mostram-se ineficientes para aplicações de *big data*.

Muitos experts e cientistas de dados definem *big data* pelas seguintes características principais (chamadas 3 Vs) (OUSSOUS et al., 2018) (GANDOMI e HAIDER, 2015):

- **Volume:** grandes volumes de dados são gerados continuamente a partir de milhares de dispositivos e aplicações (smartphones, redes sociais, sensores, logs, etc.);

- **Velocidade:** Dados são gerados de modo rápido e precisam ser processados rapidamente para que *insights* relevantes sejam extraídos;

- **Variedade:** *big data* é gerado a partir de várias fontes e em múltiplos formatos (por exemplo: documentos, vídeos, comentários, logs, etc.). Grandes conjuntos de dados são constituídos por dados estruturados e não estruturados, públicos ou privados, de origem local ou distante, compartilhados ou confidenciais, completos ou incompletos, etc.

De acordo com Gandomi e haider (2015), além dos três Vs principais, as seguintes novas dimensões foram também mencionadas como características inerentes ao *Big Data*:

- **Veracidade:** característica definida pela IBM que representa a falta de confiabilidade inerente a algumas fontes de dados. Por exemplo, os sentimentos dos clientes nas mídias sociais são de natureza incerta, pois envolvem julgamento humano. No entanto, eles contêm informações valiosas. Assim, a necessidade de lidar com dados imprecisos e incertos é outra faceta do *big data*, que é abordada usando ferramentas e análises desenvolvidas para gerenciamento e mineração de dados incertos;

- **Variabilidade (e complexidade):** novas características apresentadas pelo SAS. A variabilidade refere-se à variação nas taxas de fluxo de dados. Muitas vezes, a velocidade do *big data* não é consistente e tem picos e vales periódicos. Complexidade refere-se ao fato de que *big data* é gerado por meio de uma infinidade de fontes. Isso impõe um desafio crítico: a necessidade de conectar, combinar, limpar e transformar dados recebidos de diferentes fontes;

- **Valor:** a Oracle introduziu o Valor como um atributo definidor de big data. Com base na definição da Oracle, *big data* geralmente é caracterizado por uma "baixa densidade de valor". Ou seja, os dados recebidos na forma original costumam ter um valor baixo em relação ao seu volume. No entanto, um valor alto pode ser obtido analisando grandes volumes desses dados.

Desafios

Embora a mineração de *big data* ofereça oportunidades atrativas, pesquisadores e profissionais têm se deparado com diversos desafios ao tentarem extrair valor e conhecimento a partir desta mina de informações. As dificuldades estão em diferentes níveis, incluindo: captura de dados, armazenamento, busca, compartilhamento, análise, gerenciamento e visualização. Além disso, há problemas de segurança e privacidade, especialmente em aplicativos orientados a dados distribuídos (OUSSOUS et al., 2018).

Apesar de novas tecnologias terem sido desenvolvidas para o armazenamento de dados, os volumes de dados estão dobrando em tamanho a cada dois anos. As empresas ainda se esforçam para acompanhar a evolução de seus dados e encontrar maneiras de armazená-los com eficiência (ORACLE, 2022).

De acordo com a Oracle (2022), apenas armazenar os dados não é o suficiente. Eles devem ser usados para serem úteis, e isso depende de curadoria. Dados limpos ou relevantes para o cliente e organizados de maneira que permita uma análise significativa exigem muito trabalho. Ainda de acordo com a Oracle (2022), cientistas de dados gastam até 80 por cento de seu tempo fazendo a curadoria e preparação dos dados antes que estes possam ser utilizados.

Por fim, nota-se que a tecnologia de *big data* está mudando em ritmo acelerado. Há alguns anos, o Apache Hadoop era a tecnologia popular para esta finalidade. Em seguida, o Apache Spark foi introduzido em 2014. Hoje, uma combinação das duas estruturas parece ser a melhor abordagem.

Manter-se atualizado com a tecnologia de *big data* é um desafio contínuo (ORACLE, 2022).

Revisão #21
Criado 2022-03-08 10:34:06 UTC por FLAVIO LOPES DE MORAIS
Atualizado: 2022-12-14 09:02:06 UTC por FLAVIO LOPES DE MORAIS