BIG DATA E CIÊNCIA DE DADOS: como transformar dados em conhecimento

Rassan Dyego Romão Silva[1]

RESUMO: Até o ano 2000, podíamos dizer que a informação digital no mundo girava em torno de 25%, e ainda tínhamos muitos dados em papel, livros e outros tipos de documento. Já em meados de 2012 a 2014, o percentual de toda informação gerada que estava presente em meio digital subiu para algo em torno de 98%. Esse grande volume de dados, está explosão de dados gerados universalmente a cada instante, é chamado de Big Data, que está expondo uma nova onda de tecnologia e arquitetura destinada a extrair valor de uma imensa variedade de dados, o que permite alta velocidade com um objetivo de capturar, descobrir e analisar estas informações e dados, de forma a transforma-los em informações importantes e valiosas no âmbito de gestão de negócios. O estudo foi feito por meio de uma revisão bibliográfica, através de uma abordagem metodológica por método exploratório qualitativo.

PALAVRAS-CHAVE: Ciência de Dados. Big Data.

INTRODUÇÃO

Atualmente, estamos presenciando uma explosão de informações, seja por aplicações corporativas, seja pela internet e suas redes sociais, assim como por smartphones e celulares, que geram uma grande massa de dados complexos, estruturados e não estruturados.

Até o ano 2000, podíamos dizer que a informação digital no mundo girava em torno de 25%, e ainda tínhamos muitos dados em papel, livros e outros tipos de documento. Já em meados de 2012 a 2014, o percentual de toda informação gerada que estava presente em meio digital subiu para algo em torno de 98%. A queda de custos dos computadores e dos sistemas de armazenamento de dados e o crescimento exponencial das capacidades de processamento permitiram a disseminação da utilização de informação digital (MACHADO, 2019).

Esses dados são de extrema importância se levarmos em consideração o quanto eles podem nos auxiliar e indicar, com suas correlações, o tipo de comportamento de compra de um determinado cliente e até prever uma crise em um setor da economia, a migração de clientes para a ocorrência e surtos de doenças infeciosas, como a gripe H1N1 ou o Zika Vírus (ALECRIM, 2019).

[1] Biomédico pela Faculdade Alfredo Nasser. E-mail: rassandyego@hotmail.com.

METODOLOGIA

Foi realizado um levantamento bibliográfico, usando como descritores: ciência de dados, big data, mineração de dados, aprendizado de máquina e inteligência artificial, utilizando-se a plataforma PUBMED**,** no período de 2009 – 2019 em língua portuguesa e inglesa. Como critérios de seleção foram considerados os artigos com dados bibliográficos que abordavam ciência de dados e big data e outras informações específicas correlacionadas ao assunto. Em seguida, foi feita uma leitura analítica para ordenar as informações e identificar o objeto de estudo.

RESULTADOS E DISCUSSÃO

O mundo não apenas está cada dia mais cheio de informação como também a informação está se acumulando com muita rapidez e disseminação. A mudança de escala levou a uma mudança de estado. A mudança quantitativa gerou uma mudança qualitativa. Ciências como a astronomia e a genômica, que vivenciaram uma explosão nos anos 2000, criaram o termo "big data". Hoje o conceito está migrando para todos os campos do conhecimento humano (MACHADO, 2019).

Esse grande volume de dados, está explosão de dados gerados universalmente a cada instante, é chamado de Big Data, que está expondo uma nova onda de tecnologia e arquitetura destinada a extrair valor de uma imensa variedade de dados, o que permite alta velocidade com um objetivo de capturar, descobrir e analisar estas informações e dados, de forma a transforma-los em informações importantes e valiosas no âmbito de gestão de negócios (ALECRIM, 2019).

No ambiente hiperdigital do nosso século, quando a quantidade de dados e de novos tipos de informação é enorme, a exatidão deixa de ser objetivo principal, e sim a descoberta de tendências que essas informações podem nos oferecer (BENEVENUTO; ALMEIDA; SILVA, 2012).

Comparando com o que o Facebook armazena, acessa e analisa, pelos dados que temos, mais de 50 petabytes de informações são geradas pelos usuários, e a cada mês são gerados mais de 700 milhões de minutos por mês, assim como a cada minuto são feitos

uploads de 48 horas de vídeos no YouTube. Isso nos mostra que nunca ninguém conseguira assistir a todos os vídeos do YouTube (MACHADO, 2019).

Extremamente relevante é que, diariamente, mais de 500 milhões de mensagens são enviadas pelo twitter, com uma média de 5.7000 TPS (*tweets per second*, ou mensagens por segundo). O recorde é de 143.199 TPS. O Google processa diariamente mais de 3 bilhões de pesquisas em todo o mundo, sendo desse total 15% totalmente inéditas. Seu motor de pesquisa rastreia 20 bilhões diariamente, armazenando 100 petabytes de informação (ALECRIM, 2019).

Os bancos de dados atuais e mais usados são os bancos de dados relacionais, definidos de forma esparsa e rígida e preparados para responder a questões simples para as quais foram preparados e somente a elas com eficiência e exatidão (GRUS, 2016).

A linguagem comum para acessar esses dados armazenados em bancos de dados relacionais é a SQL, ou linguagem de consulta estruturada. Seu próprio nome evoca sua rigidez, estruturada (AMO, 2010).

A grande mudança tem ocorrido em função das estruturas não rígidas e de dados não estruturados coletados e armazenados em Big Data, sem o formato de listas ou tabelas da qual surge a linguagem NoSQL, que não exige uma estrutura predefinida para funcionar, aceitando dados de vários tipos e tamanhos para explorar (PONDEY; CHAWLA; POON; 2009).

A análise das informações contidas nesses volumes de dados é realizada por consultas diretas com acesso a dados específicos para obter as informações necessárias. Porém, essas consultas apresentam-se mais eficazes em pequenas bases de dados, de modo que quando maior a sobrecarga de informação, mais complexo sua exploração. Surgiu assim, a necessidade de uma nova geração de técnicas e ferramentas com a habilidade de auxiliar os analistas humanos de uma forma “inteligente” e “automática” na procura de informações potencialmente úteis, previamente desconhecidas nos dados (ALECRIM, 2019).

As ferramentas e técnicas empregadas para análise automática e inteligente destes grandes repositórios são os objetos tratados por uma nova área denominada Descoberta de Conhecimento em Bancos de Dados (DCBD), da expressão em inglês *Knowledge Discovery in Databases* (KDD) (GRUS, 2016).

O processo de KDD é um conjunto de atividades contínuas para o compartilhamento de conhecimento descoberto a partir de bases de dados, esse conjunto é dividido em seis etapas (AMO, 2010).

Podem-se distribuir os passos do processo de KDD, em três etapas primordiais: o Pré-processamento, a Mineração de Dados e o Pós-processamento (GRUS, 2016).

A etapa de pré-processamento é uma das mais demoradas podendo tomar até 80% de todo tempo necessário para o processo completo, devido às muito conhecidas dificuldades de integração de bases de dados heterogêneas. Nesta são realizados os seguintes passos: definição dos objetivos, coleta de dados, limpeza e pré-processamento dos dados e transformação de dados (GRUS, 2016).

A Extração de Padrões é direcionada ao cumprimento dos objetivos definidos na identificação do problema. Esta etapa é um processo iterativo, podendo ser necessária sua execução por diversas vezes até que se ajuste ao conjunto de parâmetros visando à obtenção de resultados mais adequados aos objetivos preestabelecidos. Cabe ressaltar também que, esta compreende a escolha da tarefa de MD a ser empregada, a escolha do algoritmo e a extração dos padrões propriamente dita (TAURION, 2013).

A fase de pós-processamento envolve a visualização, análise e a interpretação do modelo de conhecimento gerado pela etapa de MD. Em geral, é nesta etapa que o especialista em KDD e o especialista no domínio da aplicação avaliam os resultados obtidos e definem novas alternativas de investigação dos dados (AMO, 2010).

A Ciência de Dados ou Data Science é a disciplina que combina ideias da Estatística e da Ciência da Computação para resolver o problema da descoberta de conhecimento em bases dados. A estatística tem o papel de fornece as ferramentas para descrever, analisar, resumir, interpretar e realizar inferências sobre os dados. Por sua vez, a Ciência da Computação preocupa-se em oferecer tecnologias eficientes para o armazenamento, acesso, integração e transformação dos dados. Ou seja, o papel da Ciência da Computação é tornar viável a análise de bases de dados, muitas vezes complexas e volumosas, através de processos estatísticos. Dentre as diferentes tecnologias utilizadas para computação científica, Python é, sem dúvida, uma das que alcançou maior destaque. Trata-se de uma linguagem de programação livre, extremamente versátil e poderosa, que tem sido largamente adotada em projetos relacionados à ciência de dados (GRUS, 2016).

CONCLUSÕES

Conclui-se que, Big Data é uma coleção de conjuntos de dados, grandes e complexos, que não podem ser processados por bancos de dados ou aplicações de processamento tradicionais. Dados são transformados em informação, que precisa ser colocada em contexto para que possa fazer sentido. A Ciência de Dados utiliza os métodos estatísticos para explorar e analisar dados, fazer inferências e buscar padrões em meio de incertezas, tudo isso em novas abordagens e com o auxílio da Ciência da Computação.

REFERÊNCIAS

ALECRIM, E. **O que é Big Data?** 13 jan. 2015. Disponível em: <http://www.inforwester.com/bigdata.php>. Acesso em: 16 jun. 2019.

AMO, S. Curso de Data Mining – Aula 2 – Mineração de Regras de Associação – O Algoritmo **APRIORI**. 2010.

BENEVENUTO, F.; ALMEIDA, J. M.; SILVA, A. S. **Coleta e análise de grandes bases de dados de redes sociais online.** Capitulo 2, 2012. Disponível em: <http://homepages.dcc.ufmg.br/fabricio/download,2012.pdf>. Acesso em: 20 jun. 2019.

GRUS, J. **Data Science do Zero**. Rio de Janeiro: Altas Books, 2016.

MACHADO, F. N. R. **Big Data:** o futuro dos dados e aplicações. São Paulo: Erica, 2018.

PANDEY, S. *et al*. ***"Association rules network:** Definition and applications", in Statistical Analysis and Data Mining*. Wiley, 2009, p. 260-79.

TAURION, C. *Big Data*. **Barsport** [locais do kendle 18-19], 2013. E-book/kendle.116 p.