UNIVERSIDADE DE BRASÍLIA

Faculdade do Gama

Sistemas de Banco de Dados 2

**Trabalho Final de Banco de Dados (TF-BD)**

# Big Data

**João Vítor Morandi Lemos**

**160010195**

**Filipe Toyoshima Silva**

**160049971**

Brasília, DF

2018

**a) Introdução**

Com o aumento na quantidade de dados contidas nos bancos de dados, os sistemas gerenciadores existentes não conseguiam mais ser efetivos, o que forçou o surgimento de uma nova tecnologia: a Big Data. Esse tipo de tecnologia já vinha sendo desenvolvido há algum tempo, mas o termo surgiu e se popularizou apenas nos anos 90 [2].

O objetivo do Big Data pode ser conceituado como o gerenciamento e a análise de uma quantidade massiva de dados em um curto espaço de tempo. Essa tecnologia também pode trabalhar com dados não-estruturados, ou seja, ela pode trabalhar com imagens, geolocalização e até com posts no Facebook [1].

Sobre as características do Big Data, a maioria dos autores define essa tecnologia a partir de 5 V´s:

**Volume:** bastante auto-explicativo, se refere a grande quantidade de dados com a qual essa tecnologia trabalha. Extremamente importante para análises confiáveis dos dados, uma vez que quantos mais dados se tem disponível para treinar algoritmos de aprendizagem de máquina, por exemplo, mais confiabilidade se tem nas predições desse algoritmo. A ciência de dados se torna cada vez mais acurada com uma maior quantidade de dados disponíveis.

**Variedade:** se refere aos diferentes tipos de dados armazenados nesse tipo de banco de dados, sendo que a grande maioria deles é composta de dados não-estruturados. Trabalhar com uma variedade de dados pode não ser fácil, e quanto melhor a adaptabilidade da tecnologia de Big Data responsável pelo tratamento desses dados, mais eficiência e possibilidades são empregadas ao seus usos.

**Velocidade:** uma grande virtude do Big Data é poder realizar todo esse processamento de dados em tempo real, pois ele é utilizado em sistemas que exigem tal comportamento, pois interagem diretamente com usuários que estão acostumados com respostas rápidas de programas.

**Veracidade:** a veracidade é importantíssima para o Big Data. Se os dados contidos na base estiverem incorretos ou forem enganosos, seria preferível nem tê-los em primeiro lugar[3].

**Valor:** representa o benefício, tanto de eficiência quanto monetário, que esse tipo de tecnologia irá provocar a empresa.

**b) Principais vantagens**

Com a evolução das tecnologias de informação, cada vez mais tem se utilizado formas de coletar informações do mundo real, informações estas que estão se acumulando em centros de dados que crescem exponencialmente. Tão grande quanto o volume de dados armazenado são os problemas em se analisar tamanha quantidade de informações. Não se pode, porém, ignorar as vantagens de se fazer esse tipo de análise.

A análise de informações acerca do que tem acontecido no mundo pode oferecer uma vantagem competitiva decisiva no cenário comercial atual. Muito tem se falado a respeito de Data-Driven Strategies, ou estratégias orientadas a dados, para as quais é necessário uma inteligência no que se refere à ciência de dados e, obviamente, uma grande quantidade de dados. Esta inteligência é crucial para que se possa ter uma noção cada vez mais acurada do que pode acontecer no futuro [6], capacidade esta que se torna indispensável uma vez que as incertezas pairam sobre as decisões de grandes companhias.

Justamente para lidar com este grande volume de dados surgem as tecnologias de Big Data, que se propõem a mitigar os problemas causados por quantidades massivas de informação através de uma gestão estratégica da movimentação de dados. Cerca de 90% dos dados existentes não estão estruturados [5], o que traria outra dificuldade na análise, e por isso as estratégias de Big Data também vem a se tornar capazes de trabalhar com este tipo de dado.

Além de detectar fraudes e erros mais rapidamente, realizar operações em tempo real permite também ao banco de dados que utiliza Big Data efetividade em áreas específicas que outras tecnologias não conseguem. Algumas vantagens específicas são encontrados nas áreas: serviços financeiros, varejo, saúde, setor público, ensino e manufatura [5].

**c) Principais desvantagens**

Um problema muito comum decorrente da implementação do Big Data é a logística. Como os dados obtidos por meio dessa tecnologia são obtidos de maneira constante ao invés de periodicamente, como ocorre em outros tipos de bancos, a empresa precisaria mudar completamente sua estratégia de gerenciamento, o que custaria bastante para ela.

Outra desvantagem é o fato do Big Data ser uma tecnologia de risco, ou seja, se a empresa que deseja utilizá-la falhar em conduzir as sofisticadas análises que essa tecnologia realiza, todo o gerenciamento do banco de dados estará comprometido.

Por último, existe um problema decorrente da análise em tempo real: a privacidade. Advogados e políticos atacam constantemente equipamentos que utilizam de Big Data para obter informações pessoais de seus usuários sem autorização [6]. Uma série de notícias circulou pelos canais de comunicação no início de 2018 relatando o escândalo envolvendo a rede social Facebook e a empresa de análise de dados Cambridge Analytica [15]. A alegação seria a de que os dados dos usuários teriam sido vazados sem seu consentimento, o que poderia, inclusive, ter forte influência nas eleições dos Estados Unidos do ano de 2016, já que, através dos dados coletados, os a campanha eleitoral poderiam realizar propagandas mais direcionadas a cada tipo de eleitor.

**d) Exemplos reais**

Segundo reportagem da Software & Solutions, empresa que atua há muito tempo no campo de análise de dados, várias universidades dos Estados Unidos têm usado tecnologias de análise de quantidades massivas de dados para orientar suas decisões. A University of Alabama, por exemplo, conta com dados de mais de 38000 estudantes, os quais eram tratados manualmente no passado, o que era tedioso e consumia muito tempo [7].

Um exemplo de ferramenta no mercado capaz de auxiliar com os problemas de lidar com uma massiva base de dados é a AWS [8]. Esta ferramenta, de propriedade da Amazon, lida com tarefas normalmente trabalhosas quando se trata de volumes massivos de dados, como a movimentação de dados, o armazenamento, a análise e aplicação de Machine Learning. Um dos serviços disponíveis na AWS é a framework Amazon EMR, que suporta 19 estruturas open source para processamento de Big Data, sendo algumas delas Hadoop, Apache Spark, HBase e Presto [9]. Outra ferramenta no mercado é a Google BigQuery, integrada com a Google Cloud, oferece serviços de armazenagem em grandes volumes e processamento de dados sob demanda, capaz, ainda, de realizar algoritmos de aprendizagem de máquina nos dados armazenados [13].

Em matéria no site Big Data Business, da Hekima, uma empresa que trabalha com Machine Learning e análises de dados em geral, é relatado que o Spotify, atualmente um dos serviços de streaming de áudio mais populares do mundo, tem mostrado excelentes resultados no que se refere à análise de quantidades massivas de dados. Montando sua Social Data com informações de uso de seus milhões de usuários, tem sido possível realizar o que se chama de Spotify Audience Targeting, que se trata de entender cada vez mais o usuário, para que se

possa sugerir faixas musicais com ênfase nos gostos de cada um e de seus semelhantes [10]. A Forbes também reconhece os feitos do famoso serviço de stream em relação ao uso de informações. Segundo a revista, não há dúvidas de que o Spotify é considerado uma companhia orientada a dados [14]. E os dados não são utilizados apenas para oferecer um serviço mais personalizado aos ouvintes de música, mas também para os artistas, possibilitando que eles também possam administrar sua imagem de maneira orientada a dados coletados de seus ouvintes assíduos.

O Walmart também consegue extrair valor dos grandes volumes de dados. Bernard Marr conta em seu livro "BIG DATA, using smart big data analytics and metrics to make better decisions and improve performance", conta que a gigante de vendas lida com cerca de 2.5 petabytes de informações, gerados por milhões de transações diariamente. Além dessa, o livro também cita Google, Facebook e a já citada aqui Amazon [11].

O governo brasileiro disponibiliza o portal de dados de abertos [12], onde é possível realizar consultas sobre dados a respeito de vários ministérios e universidades. Como pode ser percebido nas aplicações citadas pelas empresas acima, com o volume de dados que o governo armazena, é possível realizar análises e inferências que auxiliem na gestão dos assuntos públicos.

**e) Referências bibliográficas**

[1]- O QUE É BIG DATA?. Canaltech. Disponível em: <https://canaltech.com.br/big-data/o-que-e-big-data/>. Acesso em: 04/11/2018

[2]- BIG DATA. Wikipedia. Disponível em: <https://en.wikipedia.org/wiki/Big_data>. Acesso em: 04/11/2018

[3]- BIG DATA´S FOURTH V. Spotless data. Disponível em: <https://spotlessdata.com/blog/big-datas-fourth-v>. Acesso em: 04/11/2018

[4]- VANTAGENS E DESAFIOS DO BIG DATA. Ipdt blog. Disponível em: <https://ipdtblog.wordpress.com/2017/04/20/vantagens-e-desafios-do-big-data/>. Acesso em: 04/11/2018

[5]- CORDEIRO, Cristiano. *Vantagens gerais e específicas do Big Data: mostramos tudo aqui!*. Disponível em: <http://www.neomind.com.br:81/blog/big-data-quais-as-vantagens-gerais-e-especificas/>. Acesso em: 04/11/2018

[6]- PROS AND CONS OF BIG DATA. Ciklum. Disponível em: <https://www.ciklum.com/blog/pros-and-cons-of-big-data/>. Acesso em: 04/11/2018

[7]- Big data in education. SAS. Disponível em: <https://www.sas.com/en_us/insights/articles/analytics/big-data-in-education.html>. Acesso em 05/11/2018.

[8]- Data Lakes and Analytics on AWS. Disponível em: <https://aws.amazon.com/pt/big-data/datalakes-and-analytics/>. Acesso em 05/11/2018

[9]- Amazon EMR. Disponível em: <https://aws.amazon.com/pt/emr/>. Acesso em 05/11/2018

[10]- Big Data Business, Hekima. Disponível em: <http://www.bigdatabusiness.com.br/spotify-e-big-data-dados-sao-musica-para-seus-ouvidos/ >. Acesso em 06/11/2018

[11]- MARR, Bernard. Big data : using smart big data, analytics and metrics to make better decisions and improve performance. ISBN 978-1-118-96583-2 (pbk.)

[12]- Portal Brasileiro de Dados Abertos. Disponível em: <http://dados.gov.br/>. Acesso em 06/11/2018.

[13]- Google BigQuery. Disponível em <https://cloud.google.com/bigquery/ >. Acesso em 06/11/2018.

[14]- MARR, Bernard - The Amazing Ways Spotify Uses Big Data, AI And Machine Learning To Drive Business Success, Forbes. Disponível em <https://www.forbes.com/sites/bernardmarr/2017/10/30/the-amazing-ways-spotify-uses-big-data-ai-and-machine-learning-to-drive-business-success/#408e68c74bd2>. Acesso em 06/11/2018.

[15]- Facebook–Cambridge Analytica data scandal - Wikipedia. Disponível em <https://en.wikipedia.org/wiki/Facebook%E2%80%93Cambridge_Analytica_data_scandal>. Acesso em 06/11/2018

**f) Base de dados**

A base de dados escolhida para se trabalhar conta com os registros de matrícula do PRONATEC de todo o Brasil. Foram retirados do portal brasileiro de dados abertos, mais especificamente do Ministério da Educação.

Notam-se alguns erros na implementação da base de dados, mas para preservação da integridade da base escolhida, formulou-se um DER fiel ao formato das informações encontradas no portal, respeitando, também, os nomes dos atributos nos arquivos baixados do site.

Os dados podem ser encontrados aqui: <http://dados.gov.br/dataset/mec-pronatec-eptc>. Foram acessados no dia 06/11/2018.

**DE-R**