REVISTA DE COMPUTAÇÃO E SISTEMAS (RCS)

ISSN XXXX-XXXX

VOLUME 1 – NÚMERO 1 – ABRIL/2016

http://www.revistacomputacaoesistemas.net

RISCOS ADVINDOS DA UTILIZAÇÃO DE BIG DATA E COMPUTATIONAL SOCIAL SCIENCE

Vivaldo José Breternitz, Leandro Augusto da Silva, Fábio Silva Lopes
Faculdade de Computação e Informática, Universidade Presbiteriana Mackenzie
Consolação, 930 – 01.302-090 – São Paulo – SP – Brazil
vjbreternitz@mackenzie.br, leandro.augusto@mackenzie.br, flopes@mackenzie.br

Abstract. *The abundance of data and the speed at which they are generated have led to changes in planning and operation in various organizational instances. Big Data, the name given to a set of technology trends that allows a new approach to the treatment and exploration of large sets of data for decision making, allows the dynamics of a society can be analyzed from the perspective of information. Computational Social Science (CSS), as this type of analysis is defined, suggests a discussion of the risks in the discovery of information in this social context. It is in this discussion that the work fits, presenting Big Data and CSS, and discussing the risks inherent in its practical uses.*

Resumo. *A abundância de dados e a velocidade em que eles são gerados têm provocado mudanças de planejamento e operação em diversas instâncias organizacionais. Big Data, nome dado a um conjunto de tendências tecnológicas que permite uma nova abordagem para o tratamento e exploração de grandes conjuntos de dados para fins de tomada de decisões, permite que a dinâmica de uma sociedade possa ser analisada, sob a perspectiva da informação. Computational Social Science (CSS), como este tipo de análise é definida, sugere uma discussão sobre os riscos na descoberta de informações nesse contexto social. É nesta discussão que o trabalho se insere, apresentando Big Data e CSS, e também discutindo os riscos inerentes em seu uso prático.*

1. Introdução

Quando uma nova tecnologia começa a emergir com força, as comunidades acadêmica e de negócios, seguidas pelo restante da sociedade, procuram conhecer os benefícios que a mesma pode trazer às pessoas e organizações. Apenas mais tarde, quando essa nova tecnologia já está sendo utilizada, os riscos decorrentes de seu uso começam a ser percebidos, partindo-se daí para a busca de soluções que possam mitigar esses riscos.

Big Data vive uma situação como a acima descrita: o entusiasmo despertado pelos benefícios que podem ser obtidos com seu uso está chegando às comunidades acadêmica e de negócios, mas quase nada vem sendo pensando em termos de riscos que podem ser trazidos pelo seu uso intensivo.

São exceções algumas considerações acerca de ameaças à privacidade: a revista The Economist [Economist, 2010] já mencionava o *trade-off* entre os potenciais ganhos econômicos trazidos pela utilização de Big Data e os riscos à privacidade. Tene e Polonetsky [2012] alertam para os riscos cada vez maiores e para a necessidade de aperfeiçoamento do marco legal relativo ao assunto; Rossouw [2012] relata propostas da Comissão Europeia (CE) no sentido de adequar a legislação da União Europeia a essa nova realidade; a CE considera, por exemplo, as normas de privacidade adotadas em março de 2012 pelo Google como "altamente arriscadas", ainda que não as tenha declarado ilegais. A visão da CE é de que unificar os dados de usuários de diferentes serviços representa risco severo para a privacidade individual.

Além dos aspectos ligados ao risco, a sociedade como um todo desconhece os conceitos básicos relativos ao Big Data.

2. Objetivos e aspectos metodológicos

Dado esse cenário, decidiu-se desenvolver este ensaio, que teve como objetivo apresentar os conceitos básicos relativos a Big Data e discutir alguns aspectos relevantes relativos aos riscos trazidos pela disseminação da utilização dessa ferramenta, especialmente aqueles ligados à Computational Social Science (CSS), de forma a gerar subsídios para os envolvidos com o tema.

Do ponto de vista metodológico, o ensaio foi produzido a partir de pesquisa de natureza exploratória, que conforme dizem Selltiz *et al* [2001], tem como objetivo proporcionar maior familiaridade com o problema, torná-lo mais explícito e construir hipóteses para posterior investigação.

À pesquisa exploratória somou-se a experiência profissional e acadêmica de seus autores, gerando o ensaio, que Ortega y Gasset [2004] define como “ciência sem prova explicita”, qualificando-o como um texto literário breve, que expõe ideias, críticas e reflexões a respeito de um dado tema, defendendo um ponto de vista pessoal e subjetivo sobre o mesmo sem se pautar por formalidades como documentos e provas empíricas ou dedutivas de caráter científico.

Passa-se agora a apresentar os conceitos básicos e a discutir aspectos relevantes relativos ao assunto, conforme os objetivos acima mencionados.

3. Apresentando Big Data e tecnologia a ele associada

Ainda não há uma definição precisa para Big Data, mas pode-se usar o termo para designar um conjunto de tendências tecnológicas que permite uma nova abordagem para o tratamento e exploração de grandes conjuntos de dados para fins de tomada de decisões.

Alguns autores, como Zikipoulos *et al* [2012] dizem que Big Data se caracteriza por quatro aspectos: volume, velocidade, variedade e veracidade.

O aspecto volume refere-se ao fato de que a quantidade de dados disponível em forma digital cresce de maneira exponencial, provenientes não só de sistemas convencionais, também de fontes como Facebook, Tweeter, You Tube, RFID, eletrônica embarcada, telefones celulares e assemelhados, sensores de diversos tipos etc.

Ao final de 2012, McAfee e Brynjolfsson [2012] estimavam que cerca de 2,5 *exabytes* de dados eram criados a cada dia, e que este número irá dobrar a cada 40 meses, aproximadamente. Os mesmos autores dizem que na atualidade a cada segundo, mais dados transitam pela internet do que o total armazenado na mesma há 20 anos. Apenas o Walmart coleta mais de 2,5 *petabytes* a cada hora, derivados das transações efetuadas por seus clientes.

McAfee e Brynjolfsson [2012] apresentam outro aspecto relevante de Big Data: a velocidade em que dados podem ser capturados e processados, quase em *real time*, podendo dar a uma organização vantagem competitiva. Exemplificam essa afirmação relatando experimento conduzido pelo grupo de pesquisa do Professor Alex Pentland, do MIT Media Lab: o grupo capturou dados relativos à localização de celulares de forma a inferir quantas pessoas colocaram seus carros nos estacionamentos de lojas do grupo americano Macy's no Black Friday de 2011 (data que marca o inicio da temporada de compras de Natal nos Estados Unidos); isso permitiu estimar com precisão as vendas dessas lojas antes mesmo que elas ocorressem, gerando vantagens competitivas que poderiam ser utilizadas por áreas comerciais, de *marketing* e por terceiros, como investidores em bolsas de valores.

No que se refere à variedade, cabe registrar que além de fontes diferentes, tais dados têm, frequentemente, características que fogem das tratadas pelos sistemas convencionais, não sendo estruturados e referindo-se a coisas como movimento, temperatura, umidade e até mesmo variações na composição química do ar (Lohr, 2012]. Neste aspecto, a internet das coisas (Internet of Things), é uma nova promessa de integração de várias tecnologias e soluções de comunicação, de modo a distribuir inteligência para diferentes dispositivos de modo a prover interação e cooperação entre eles (Atzori, Iera e Morabito, 2010]. No entanto, além da interação e cooperação, estes dispositivos também geram dados que podem ser armazenados, compartilhados, agregados e analisados.

O aspecto veracidade está relacionado ao fato de que os dados não são "perfeitos", no sentido de que é preciso considerar o quão bons devem ser os mesmos para que gerem informações úteis e também os custos para torna-los bons.

Alguns autores consideram um quinto aspecto, a validade dos dados, ou seja, sua vida útil, o tempo em que os mesmos precisam ser mantidos [Taube, 2012]. Esses aspectos são coletivamente chamados 4V ou 5V.

As ferramentas computacionais, por outro lado, vêm acompanhando o crescimento dessa velocidade e do volume de dados, em termos de capacidade de armazenamento e processamento. Destacam-se nesse assunto as pesquisas em corrente contínua de dados (*stream computing*) e em técnicas de inteligência artificial (*artificial intelligence*).

No modelo convencional de armazenamento de dados e tomada de decisão, a organização filtra dados dos seus vários sistemas e após criar um *data warehouse*, constroem-se consultas (*queries*) que subsidiarão a tomada de decisões. Na prática faz-

se garimpagem em uma base de dados estática, que não reflete o momento, mas sim o contexto de horas, dias ou mesmo semanas atrás. Com *stream computing*, esse *mining* ocorre em tempo real, com uma corrente contínua de dados *(streaming data*) atravessando um conjunto de *queries* - isso pode ser considerado um novo paradigma.

Na Inteligência Artificial, por sua vez, destacam-se os estudos em processamento de linguagem natural (*natural-language processing*), reconhecimento de padrões (*pattern recognition*) e aprendizado de máquina (*machine learning*) que podem ajudar a extrair dos grandes volumes de dados (estruturados ou não estruturados) conhecimento para auxiliar a tomada de decisões [Lohr, 2012].

Observa-se também a evolução relativa aos modelos de persistência de dados. O modelo hierárquico utilizado nos anos 1960 e 1970 foi substituído pelo modelo relacional nos anos 1980, e este se mantém em uso na grande maioria das aplicações em produção. Contudo, a nova geração de bancos de dados conhecidos como NoSQL, apresentam-se como novas opções mais eficientes para manipulação de grandes volumes de dados não estruturados, principalmente abordando pontos como estrutura não relacional, plataforma distribuída, código aberto e horizontalmente escaláveis [NoSQL, 2013].

Armazenar é apenas parte do negócio. A recuperação e análise dos dados têm ganho atenção no que diz respeito ao desenvolvimento de ferramentas que ampliam a capacidade analítica em grandes volumes de dados. Contudo, as ferramentas de *Business Intelligence* (BI), que tem esse objetivo, não trabalham em tempo real. Esta característica está atribuída às novas ferramentas denominadas *Business Analytics* (BA), conforme diz Gnatovich [2006].

As possibilidades de aplicação desses conceitos são inúmeras, em finanças, saúde segurança, manufatura etc. McAfee e Brynjolfsson [2012] conduziram estudos que levaram à conclusão de que as empresas que efetivamente utilizam Big Data são 5% mais produtivas e 6% mais lucrativas que seus competidores – na atualidade esses números são um poderoso argumento em prol da utilização dessa abordagem. Ben Waber, pesquisador do Instituto de Tecnologia de Massachusetts (MIT), afirma: "Se as pessoas aprenderam alguma coisa nas últimas décadas, foi que usar dados para construir organizações é melhor do que seguir instintos." [Battibugli, 2013].

Moraes [2012] relata como a aplicação de Big Data ajudou na campanha de reeleição do presidente norte-americano, Barack Obama, permitindo orientar voluntários, indicar as melhores formas de arrecadar fundos e apontar quem poderia ser convencido a apoiar a reeleição do presidente; os responsáveis pela campanha deram prioridade ao uso de Big Data em detrimento da propaganda pela televisão.

Os responsáveis pela campanha usaram a Amazon Web Services para armazenar e processar o enorme volume de dados capturados. Foram adotadas ferramentas de computação em nuvem para lidar com bancos de dados, como o Amazon DynamoDB e Amazon RDS. Uma das principais preocupações foi permitir que a base de dados fosse trabalhada por diferentes aplicativos escritos em diversas linguagem de programação – para isso, se desenvolveu o Narwhal, um conjunto de serviços que funcionava como *interface* entre os dados e os muitos programas criados para a campanha.

4. Apresentando Computational Social Science (CSS)

Computational Social Science (CSS) pode ser definida como a ciência que compreende a investigação da dinâmica social conduzida de forma interdisciplinar, sob a perspectiva da informação e por meio do uso de sistemas computacionais avançados [Cioffi–Revilla, 2010], como Big Data; a academia já começa a discutir sua aplicação em estudos ligados às ciências sociais, políticas públicas e comportamento de indivíduos e grupos [Global Pulse, 2012].

King [2013] descreve como estudantes de cursos ligados às ciências sociais, usualmente em nível de pós graduação, estão passando a receber treinamento formal em computação, como parte de sua formação para que possam atuar na área; este fenômeno vem se cristalizando com a criação em algumas universidades de departamentos ou cursos usualmente chamados "Computational Social Science" ou "Applied Computational Science"; na Harvard University foi criado o IQSS, Institute for Quantitative Social Sciences (http://iq.harvard.edu), com o objetivo de dar suporte à pesquisa na área.

Este novo campo de conhecimento é impulsionado pelos fatores anteriormente mencionados como capacidade computacional, ferramentas de apoio e o grande volume de dados gerado por diversos dispositivos e acumulados em diferentes repositórios.

5. A utilização de Computational Social Science - riscos

A capacidade, inerente a Big Data, de coletar e analisar grandes volumes de dados permite que se revele padrões referentes a indivíduos e grupos e que se simule o comportamento dos mesmos quando alteradas determinadas variáveis [Agarwal *et al*, 2008].

A sociedade como um todo deve preocupar-se com a utilização dessas capacidades, por empresas, governos e outros tipos de organização – empresas e governos são, na atualidade, os detentores de vastas quantidades de dados que podem ser utilizados para CSS.

O senso comum diz que o acesso a certos dados não deve ser permitido sem muitos cuidados; por essa razão existem leis que restringem a gravação de conversas telefônicas e o acesso a registros médicos, garantem a inviolabilidade da correspondência etc.

A gravidade dos danos gerados por acesso e uso indevidos é quase sempre óbvia. No entanto, a combinação de dados aparentemente inofensivos provenientes de diversas bases ou a análise de grandes bases de dados pode gerar informações potencialmente perigosas para indivíduos, organizações e até mesmo estados, e isso é difícil de prever com suficiente antecipação. A falta de transparência acerca da forma com que dados são agregados e analisados, combinada com a dificuldade em se prever quais informações podem vir a se tornar perigosas, leva a situações em que indivíduos (e mesmo organizações) tenham pouca percepção acerca dos efeitos potencialmente deletérios do avanço do uso da CSS.

A aplicação de CSS pode ser voltada a coisas aparentemente inofensivas, como propaganda de produtos de consumo, mas, quando aplicada às áreas de antropologia social e ciência política, pode ser usada para atentar contra a democracia. Combinando essas aplicações, pode-se chegar a cenários em que CSS pode ser utilizada para dirigir propaganda política (não apenas eleitoral), selecionar quais informações podem chegar a determinados grupos, quais não podem etc. Isso torna-se particularmente grave

quando está ficando claro que as informações geradas estarão acessíveis aos governos, àqueles capazes de pagar por elas ou às empresas que conseguem reunir grandes volumes de dados, cujos exemplos clássicos são Facebook e Google [Oboler *et al*, 2012].

Conforme os atuais Termos de Serviço do Google, implementados em 2012 e aos quais seus usuários não tem a opção de se furtarem [Google, 2012], a empresa pode avaliar seus usuários, combinando e analisando dados coletados em todos os seus serviços, como por exemplo, histórico de pesquisa, utilização do Google+, conexões a redes sociais, uso do Gmail, compras *online* pagas com Google Wallet, arquivos de fotos etc. – as práticas do Google acerca de privacidade são frequentemente questionadas [Estado, 2013].

Com base nessa avaliação, os usuários poderiam ser direcionados para vídeos do You Tube e notícias do Google News (também vídeos e notícias poderiam ser "ocultados" do usuário) de forma a mostrar que um determinado partido político concorda com as opiniões desse usuário (ou que outro não concorda), influenciando sua maneira de pensar, por exemplo.

Oboler (2012), diz que tais direcionamentos poderiam ser feitos de maneira sutil, pois a CSS permite prever a eficácia de mensagens diferentes para pessoas diferentes – assim mensagens inócuas poderiam ser encaminhadas em meio a outras mais "agressivas", de forma a não caracterizar uma atuação parcial. Esses serviços poderiam ser usados não só por uma empresa como o Google, mas também vendidos a terceiros, influenciando resultados de eleições, o que é antidemocrático. Eleições envolvem frequentemente dezenas ou mesmo centenas de milhões de pessoas, não sendo, por essa razão, surpresa o fato de que elas estão entre os mais estudados fenômenos sociais e consequentemente um campo naturalmente candidato à aplicação de CSS [Fortunato e Castellano, 2012].

Jones *et al* [2013] concluíram ser possível estabelecer a força da ligação entre pessoas simplesmente analisando suas interações via Facebook, sem considerar os atributos dessas pessoas. Identificando ligações fortes, é possível também influenciar comportamentos, desde os relativamente inócuos como recomendação de produtos como àqueles ligados à política.

A aplicação de CSS, para o bem ou para o mal, é limitada pela disponibilidade de dados. Temas ligados à captura de dados a serem utilizados por aplicações ligadas à CSS e sobre o acesso às informações por eles geradas devem passar a ser consideradas questões chave em termos de políticas públicas.

Ao limitar a aquisição, compartilhamento e uso de dados, e pela sensibilização da sociedade para as implicações de sua disponibilidade, especialmente em termos éticos, é possível limitar os problemas aqui relatados, hoje ainda relativamente raros, mas que tendem a aumentar exponencialmente se nada for feito pela sociedade. O uso da CSS pode ser um grande benefício da busca por conhecimento, mas como acontece em relação a todos os avanços científicos, a sociedade precisa estar consciente de seus riscos.

6. Considerações finais

Considerando as questões aqui apresentadas, observa-se que o crescimento deste campo de conhecimento é evidente, bem como, os benefícios envolvidos. Contudo, assim como em outras ciências, faz-se necessária a discussão mais ampla acerca das questões éticas e os riscos inerentes a este contexto.

Na ótica da tecnologia, o fenômeno Big Data está sendo muito positivo, pois motivou avanços na área de persistência de dados, tanto no que diz respeito a *hardware* como *software*. Assim como acontece no campo da moda, as empresas de tecnologia estão vivendo, de certa forma, uma volta ao passado, como acontecia outrora com os *mainframes*: os principais *players* estão comercializando *hardware* para BI/BA com *software* embarcado, na perspectiva de alcançar melhor *throughput* nas operações de recuperação de informação.

De modo complementar, os novos paradigmas de persistência, da família NoSQL, estão contrapondo a teoria de que um sistema gerenciador de banco de dados único é quase sempre suficiente para atender as necessidades de armazenamento e recuperação de dados de uma organização; soluções híbridas parecem melhor atender demandas em tempo de Big Data, que envolvem questões de escalabilidade em ambientes distribuídos e heterogêneos.

Assim como as questões tecnológicas evoluem, é importante evoluir na mesma cadência a discussão sobre os aspectos éticos e os riscos envolvidos.

O assunto é complexo e globalizado, na medida que a tecnologia está derrubando fronteiras geográficas. Empresas geram armazéns de dados globalizados. Mesmo que um país determine leis que regulamentem o uso de dados para CSS, tal legislação está limitada às fronteiras geográficas daquele país. Não são casuais os conflitos que a Google tem em países como China e França.

Governos a parte, a construção de perfis individuais com base em conteúdos digitais, bem como o seu uso nas diversas formas de aplicação, compõem um tema que carece de discussão mais aprofundada na sociedade.

Referências Bibliográficas

Agarwal, R., Gupta, A. K. e Kraut, R. (2008). Overview - the interplay between digital and social networks. In *Information Systems Research*, vol. 19, nº. 3.

Atzori, L., Iera, A. e Morabito, G. (2010). The Internet of Things: A survey. In *Computer Networks: The International Journal of Computer and Telecommunications Networking*, vol. 54, ed. 15.

Battibugli, A. O. (2013). Big Data antecipa a morte do currículo. In *EXAME*, edição de 26.05.2013.

Cioffi–Revilla, C. Computational Social Science. (2010). In *Wiley Interdisciplinary Reviews: Computational Statistics*, vol, 2, nº 3.

Chui, M., Löffler, M. e Roberts, R. (2010). The internet of things. In *McKinsey Quarterly*, v. 2.

Clifford, S. (2012) Retail frenzy: prices on the Web change hourly. In *The New York Times,* edição de 30.11.2012.

Economist. (2010). The data deluge. In *The Economist*, edição de 27.02. 2010.

Estado. (2013). Google Play tem privacidade questionada. In *O Estado de S. Paulo*, edição de 16.02.2013.

Fortunato, S. e Castellano, C. (2012). Physics peeks into the ballot box. In *Physics Today*, vol. 65, nº 10, 2012.

Global Pulse. (2012). Big Data for development: challenges & opportunities. Nova Iorque: Global Pulse.

Gnatovich, R. (2006). Business Intelligence versus Business Analytics - What's the difference?, disponível em http://www.cio.com/article/18095/Business_Intelligence_Versus_Business_Analytics_What_s_the_Difference, acessado em 03.07.2013.

Google. (2012). Termos de serviço do Google, disponível em http://www.google.com/intl/pt-BR/policies/terms/, acessado em 14.02.2013.

Jones, J. J., Settle, J. E., Bond, R. M., Fariss, C. J., Marlow, C. e Fowler, J. H. (2013). Inferring tie strength from online directed behavior, disponível em http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0052168, acessado em 14.02.2013.

King, G. (2013). Restructuring the Social Sciences: reflections from Harvard's Institute for Quantitative Social Science, disponível em http://gking.harvard.edu/files/gking/files/iqsss.pdf, acessado em 13.01.2013.

Lohr, S. (2012). "The Age of Big Data" http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html?_r=1&scp=1&sq=Big%20Data&st=cse, acessado em 02.01.2013.

Mcafee, A. e Brynjolfsson, E. (2012). Big Data: The Management Revolution. In *Harvard Business Review*, edição de outubro de 2012.

Moraes, M. (2012). Big Brother Obama. In *InfoExame*, edição de dezembro de 2012.

NoSQL. (2013). Your Ultimate Guide to the Non - Relational Universe, disponível em http://nosql-database.org, acessado em 05.07.2013.

Oboler, A., Welsh, K. e Cruz, L. (2012). The danger of Big Data: social media as computational social science. In *First Monday*, vol 17, nº 7.

Ortega Y Gasset, J. (2004) Meditaciones del Quijote - in: Obras Completas, vol. I. Madrid: Taurus.

Rossouw, L. (2012). Big Data – Grandes Oportunidades. In *Gen Re – Risk Insights*, vol. 16, nº 2.

Selltiz, C., Wrightsman, L. S. e Cook, S. W. (2001). Métodos de pesquisa nas relações sociais. 2. ed. São Paulo: EPU.

Taube, B. (2012). Leveraging Big Data and real-time analytics to achieve situational awareness for smart grids (white paper). Redwood City: Versant Corporation U.S. Headquarters.

Taurion, C. (2011). Big Data: nova fronteira em gerenciamento de dados, disponível em http://www.ibm.com/developerworks/mydeveloperworks/blogs/ctaurion/entry/big_data_nova_fronteira_em_gerenciamento_de_dados?lang=en, acessado em 17.01.2013.

Tene, O. e Polenetsky, J. (2012). Privacy in the age of Big Data - a time for big decisions, disponível em http://www.stanfordlawreview.org/online/privacy-paradox/big-data, acessado em 14.01.2013.

Zikopoulos, P., De Roos, D., Parasuraman, K., Deutsch, T., Giles, J. e Corrigan, D. (2012). Harness the power of Big Data - The IBM Big Data Platform. Emeryville: McGraw-Hill Osborne Media.