Revista da
ESPM

REVISTA DA ESPM ▪ ANO 20 ▪ EDIÇÃO 95 ▪ Nº 5 ▪ SETEMBRO/OUTUBRO 2014 ▪ R$ 28,00

Entrevistas

TECNOLOGIA DA INFORMAÇÃO

O mundo entra na era da computação cognitiva

Artigos

Artigos

Big data ou *big problems*?

Os mitos e as verdades gerados a partir da análise do grande volume de informações que transita hoje na internet

Por Eduardo de Rezende Francisco

"O *big data* é mais uma tacada de marketing. Propagado pelas empresas de tecnologia e de consultoria de sistemas, o termo *big data* parece mal empregado. Isso é intencional e lhes possibilita divulgar o que elas desejam, pois seus mercados não sabem exatamente o que significa *big data*." Palavras de Stephen Few, em seu blog sobre *Visual Business Intelligence*, que discutiu esse conceito, conforme minha tradução livre. Essa é, no mínimo, uma posição polêmica.

Vivemos um contexto empresarial em que todos, ou quase todos, querem "contratar soluções de *big data*". Executivos de grandes corporações ouvem falar de *big data* em eventos sobre tendências tecnológicas, comportamento do consumidor, e logo querem implementar em suas organizações. Mas será que entendemos quais são os "problemas de *big data*", para daí concluirmos se é uma necessidade premente ou futura de nossas organizações? A confusão gerada pelo termo ainda persiste – por um *gap* de significado ou motivada pela indústria, conforme Stephen Few sugere. Mas, afinal de contas, o que é *big data*?

Para começo de conversa, o termo é realmente infeliz e traduz apenas um aspecto dentre vários que chegaram à arena digital e estão mudando, de forma significativa, o panorama de negócios e a vida da sociedade em geral. *Big data* são dados cuja escala, distribuição, diversidade e/ou velocidade de criação requerem o uso de novas tecnologias de armazenamento e análise para permitir a captura do valor inserido neles. E isso é mais amplo do que o termo "big" possa abranger.

SHUTTERSTOCK

Mitos e verdades

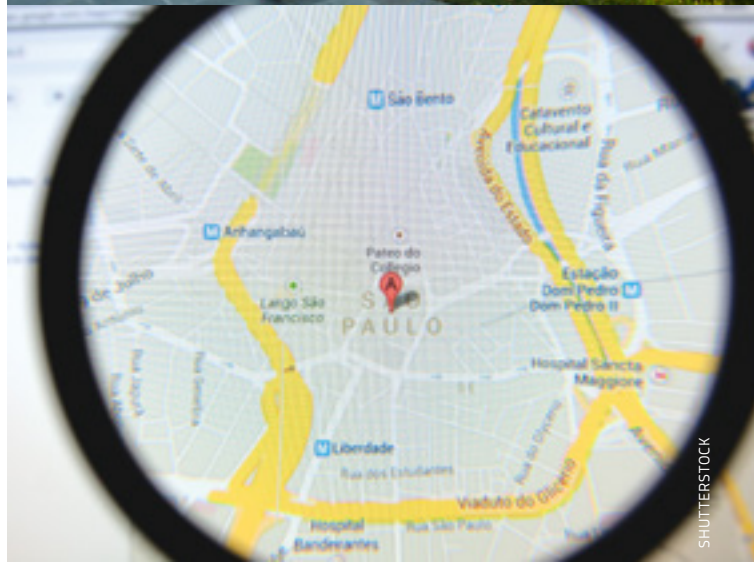
O "big" do *big data* se traduz no volume. Ordens de grandeza como a que nos acostumamos para representar o tamanho de nossos arquivos e informações na internet, como megabytes (1 milhão de bytes) ou gigabytes (1 bilhão de bytes), já não representam mais as escalas que condicionam as informações na era do *big data*. Não é raro falarmos em petabytes (1 milhão de gigabytes) ou mesmo em exabytes (mil petabytes) para representar implantações de projetos envolvendo *big data* em grandes corporações no Brasil e no mundo, voltados aos mais diversos mercados, como o de varejo supermercadista, de prospecção de minério e petróleo ou ainda para a previsão de microclima. Estima-se que em 2020 tenhamos 35,2 zettabytes (ou 35.200 exabytes) de informações no universo digital, segundo números da EMC, uma das gigantes do setor de armazenamento e processamento de dados.

Outra perspectiva do *big data* é a velocidade. Aliada ao grande volume de informações que transita atualmente na internet, hoje a velocidade de tráfego é impressionante. Um levantamento da Qmee, referente a 2013, sabe-se que em apenas um minuto a rede mundial movimenta 204 milhões de e-mails, mais de 2 milhões de consultas no google.com, 278 mil tweets, 1,8 milhão de curtidas no Facebook e 72 novas horas de vídeo são disponibilizadas no YouTube. Pouca coisa, não?

Mas a característica que diferencia a velocidade no mundo do *big data* não é seu aumento, e sim sua capacidade de ser cada vez mais assíncrona e *real time*. Em outras palavras, não conseguimos mais controlar a velocidade com que as informações trafegam. No contexto em que precisamos tomar decisões com rapidez, para que isso gere valor para as organizações, a velocidade não controlada traz um grande desafio para os modelos analíticos – eles precisam se reinventar permanentemente para se manterem úteis.

Que valor tem para uma pessoa estimar como estará o trânsito na avenida Paulista, na cidade de São Paulo, às 17 horas, se essa estimativa for obtida somente às 19 horas? Ou estimar cinco minutos depois se o uso do cartão de crédito em uma transação de compra fora do país deve realmente estar sendo realizado pelo portador do cartão ou se é uma fraude? Precisamos obter a melhor estimativa possível sobre a informação no momento em que ela está acontecendo.


LATINSTOCK



SHUTTERSTOCK

Modelos tradicionais de previsão tendem a se tornar lentos se considerarmos volumes muitíssimo grandes de dados. E é aí que o *big data analytics* surge como nova grande força nas organizações – como analisar de forma coerente e rápida informações em tempo real e em quantidade não controlada. Afrouxamos um pouco a confiança nas estimativas tradicionais para ganharmos, significativamente, em desempenho – e isso representa, realmente, um valor para as organizações.

Uma terceira perspectiva do *big data* é a variedade. De 80% a 90% dos dados da internet não são estruturados – páginas "www" em formato HTML ou XML, dados de *clickstream*, fotos, imagens, vídeos, textos em linguagem natural, mapas etc. Tipicamente, informações coletadas a todo instante por radares de trânsito, sensores de clima, câmeras de segurança, além de comentários e posts de redes sociais. Essas informações contemplam um panorama riquíssimo de significados para as empresas, que pode ser apropriado a vários contextos: avaliar como está a reputação de uma marca a partir de

LATINSTOCK

SHUTTERSTOCK

Em um minuto, a web movimenta 204 milhões de e-mails, mais de 2 milhões de consultas no google.com, 278 mil tweets e 1,8 milhão de curtidas no Facebook

comentários de usuários seguidores, analisar o perfil, interpretar de forma automática um comando de voz feito por um cliente em uma ligação no call center, inferir como está o trânsito de uma determinada via a partir de imagens sequenciais, suspeitar sobre o trajeto de uma pessoa no estacionamento de um shopping center etc.

Tudo isso complementa análises mais tradicionalmente realizadas pelas empresas, que envolvem dados estruturados: cadastro de clientes, transações comerciais, CRMs, informações bancárias, por exemplo. Dados mantidos nos tradicionais sistemas gerenciadores de bancos de dados (RDBMSs, na sigla em inglês), normalmente em modelos relacionais similares ao proposto por Edgar F. Codd, em 1969, pesquisador da IBM. Atualmente, esses dados já não dão conta de toda a necessidade informacional que um executivo em uma empresa precisa para dar segurança às suas decisões. E por isso os dados não estruturados são cada vez mais necessários.

Além destes três "Vs" (volume, velocidade e variedade), muitos adicionam, ainda, o quesito veracidade como uma das características do *big data*, que se traduz em "como realmente confiar nas informações que se utilizam para a tomada de decisão nas organizações". Mais de 90% dos executivos acreditam que o conhecimento pode ser mais bem aproveitado. Isso porque apenas 1% dos dados corporativos é usado, efetivamente, para análise nas grandes corporações. Logo, cerca de 90% das estratégias corporativas fracassam. Adiciona-se a isso o fato de que não conseguimos controlar as informações que consultamos na internet – não sabemos se há notícias falsas ou verdadeiras transitando, e como diferenciá-las.

A veracidade se confunde com o quinto "V" do *big data*: valor. Qual é o valor real que conseguimos adicionar aos negócios quando trabalhamos no contexto do *big data*? Como inferir relevância em uma vastidão de informações? Como saber, em uma busca de informações no Google, que traz 538 mil referências, quais realmente são as mais relevantes para mim? Nem sempre as mais referenciadas (que se encontram no topo "não promocional" da sua lista de respostas) são as que realmente me beneficiarão.

A tecnologia que suporta o *big data*, atualmente, deve cuidar de dois grandes processos: o que organiza o acesso a essas informações e o que organiza a análise das informações – essa também conhecida como analytics ou *big data analytics*.

Armazenar os dados do *big data* não é uma tarefa fácil – eles estão convivendo e se reinventando a todo instante na já famosa *cloud*. E talvez não seja muito inteligente, ou pelo menos é altamente custoso tentar replicá-lo estruturalmente em algum outro lugar da rede mundial. Como os peta, exa e zettabytes de dados são tipicamente não estruturados, os bancos de dados batizados de NoSQL (contração de "*Not only SQL*") permitiram que os dados fossem organizados ou pelo menos referenciados.

O artifício técnico que permitiu que os dados continuassem onde estão e fossem rapidamente consultados foi o algoritmo denominado MapReduce, que baseia, atualmente, o buscador do Google. Ele é composto de duas

grandes fases: o "Map", que dispara diversos agentes para os locais em que estão os dados (páginas da web, arquivos na internet), e o "Reduce", que inicialmente consolida os resultados totais da sua busca a partir de consolidações locais de cada agente. Isso torna o processo extremamente rápido, permitindo que consultas a milhões de páginas da internet sejam feitas em milissegundos no computador pessoal ou smartphone de qualquer pessoa.

Essa ideia não é de fato original, mas não havia sido convenientemente implementada até então. A ferramenta Hadoop é uma das que mais se popularizaram como tecnologia que suporta o *big data* a partir da implementação do MapReduce.

Organizando e analisando as informações

Os dilemas que as empresas vivem no contexto do *big data* fortalecem a necessidade de organizarmos as informações para podermos, efetivamente, analisá-las de forma razoável e factível.

Segundo a consultoria Gartner Group, entre 70% e 80% das informações relevantes nos processos decisórios têm caracterização espacial. Isso significa que questões críticas de negócio nas organizações têm, na maioria dos casos, componentes como "onde?". Onde está o meu cliente, e meus fornecedores? Para onde devo expandir minha operação? Onde está minha concorrência?

A natureza geográfica das informações possibilita a integração de dados, informações, processos, inclusive "*big data*", pelo simples fato de eles conviverem em um mesmo espaço geográfico. Essa integração induz a uma visão sistêmica (integrada, ampla, abrangente e holística) da maioria das questões necessárias às tomadas de decisão. Por exemplo: caracterizar espacialmente um tweet disparado por um cidadão sobre a qualidade dos serviços públicos da cidade. Essa informação pode ser georreferenciada, principalmente se o tweet tiver partido de um dispositivo móvel. Isso significa que a opinião dos contribuintes em geral pode variar de forma significativa conforme a região da cidade, e isso, sim, pode servir de insumo prático para a ação local do poder público.

As empresas consomem informações geográficas normalmente em processos de suporte à operação, para descrição, expansão, segmentação e otimização do território de atuação.

No contexto do *big data analytics*, para ajudar a organizar as informações, os modelos estatísticos estão se aprimorando para poderem abarcar as informações de localização e outros predicados espaciais (pertinência, conectividade e proximidade). E essa é uma tendência do *big data* analytics: o uso de técnicas de estatística espacial.

Muitos negócios podem se beneficiar dessas técnicas: modelos tradicionais de previsão de renda e de crédito, no segmento bancário e no mercado financeiro em geral; modelos de gestão de risco e desempenho, que qualificam e analisam influências locais no aumento de risco para seguradoras; e o mercado imobiliário, para a correta gestão de compra e venda de ativos e o mapeamento de regiões, visando a melhor precificação de imóveis, em um contexto em que a dinâmica urbana territorial é essencial.

O *big data* analytics tem grande potencial de investigação sob a perspectiva geográfica, pela relevância que essa natureza de informação tem e pelo fato de que dispositivos móveis (com GPS ou outros sistemas de posicionamento global) como *devices* de informação são uma realidade em regiões urbanas no mundo. O número de celulares no Brasil já ultrapassou a população brasileira – temos 272,6 milhões de assinaturas atualmente no país, o que compreende 115% da população e representa o quinto maior mercado de *mobile* no mundo.

LATINSTOCK

Dados de redes sociais, que compreendem atualmente um dos tipos de dados não estruturados mais analisados pelas empresas, podem evoluir de simples contagens de *hashtags* para *sentimental analysis*. Analisar o humor dos comentários feitos sobre uma determinada empresa, ou sobre um evento, e vincular essa análise à posição geográfica em que foram disparados. Isso tudo em tempo real, envolvendo um volume não controlado de tweets – bem-vindo ao *big data analytics*!

A internet das coisas (ou *internet of things*, *IoT*), revolução tecnológica que possibilitará a real comunicação entre aparelhos das casas das pessoas ou entre sistemas públicos dotados de sensores, potencializará ainda mais o papel do *big data analytics* no contexto do espaço geográfico e das possibilidades de captura de comportamento e informação sobre o mundo real em seus diversos processos.

O futuro que nos espera

Empresas de diversos portes podem se beneficiar do *big data* em muitas perspectivas. Mas, infelizmente, poucas são as que enveredaram pelo uso da informação geográfica em seus modelos analíticos. Esse número não chega a uma dezena delas.

A complexidade analítica traz, por um lado, o desafio da implementação técnica. No entanto, por outro, a dificuldade não é técnica. A maior barreira para o uso de *big data* pelas organizações está no aspecto cultural. Confunde-se o *big data* (na forma como foi apresentado neste artigo) com dados "big". Problemas empresariais envolvendo grandes volumes de dados já existem no "mundo real" há muitos anos. Falta às organizações a consciência de que o contexto analítico é novo quando as demais dimensões do *big data* estão presentes. As estruturas profissionais ou organizacionais que deverão resolvê-lo devem ser híbridas, formadas por profissionais de diversas áreas.

Atualmente, cunhou-se o termo data science, ou ciência de dados, como evolução da tradicional *business intelligence* para endereçar essa questão. O profissional de *data science* deve ter conhecimento de tecnologia e de negócios e ter habilidades analíticas para realmente poder encarar o *big data*. É raro essas habilidades estarem disponíveis em um mesmo profissional. Por isso, as empresas devem construir times multidisciplinares, com acesso quase que irrestrito, mas seguro, às informações, e que estejam independentes de estruturas hierárquicas rígidas, como as áreas de TI.

Bons sinais apresentam-se já no presente. A ESPM lançou recentemente o curso Sistemas de Informação em Comunicação e Gestão, que forma profissionais com habilidades em tecnologia e negócios, com uma trilha especial de *big data analytics*.

O futuro dos sistemas de informação está em quatro grandes tendências, conforme diversos *vendors* de tecnologia definem: mobilidade; *cloud*; *big data* e *analytics*; e *social business*. Isso tudo aliado a questões fundamentais, como o uso seguro e principalmente ético das informações.

O que nos resta: acompanhar estas tendências ou liderar seu entendimento e melhor aplicar em contextos de decisão? Certamente a segunda opção é a que fará diferença para o futuro. De *big data* para *big problems*. E que venham os *data scientist*!

Eduardo de Rezende Francisco
Cientista da computação, professor do NDE do curso de sistemas de informação em comunicação e gestão da ESPM e do curso de administração de empresas da FGV-Eaesp, consultor em geoinformação e sócio-fundador da Meia Bandeirada e do GisBIw