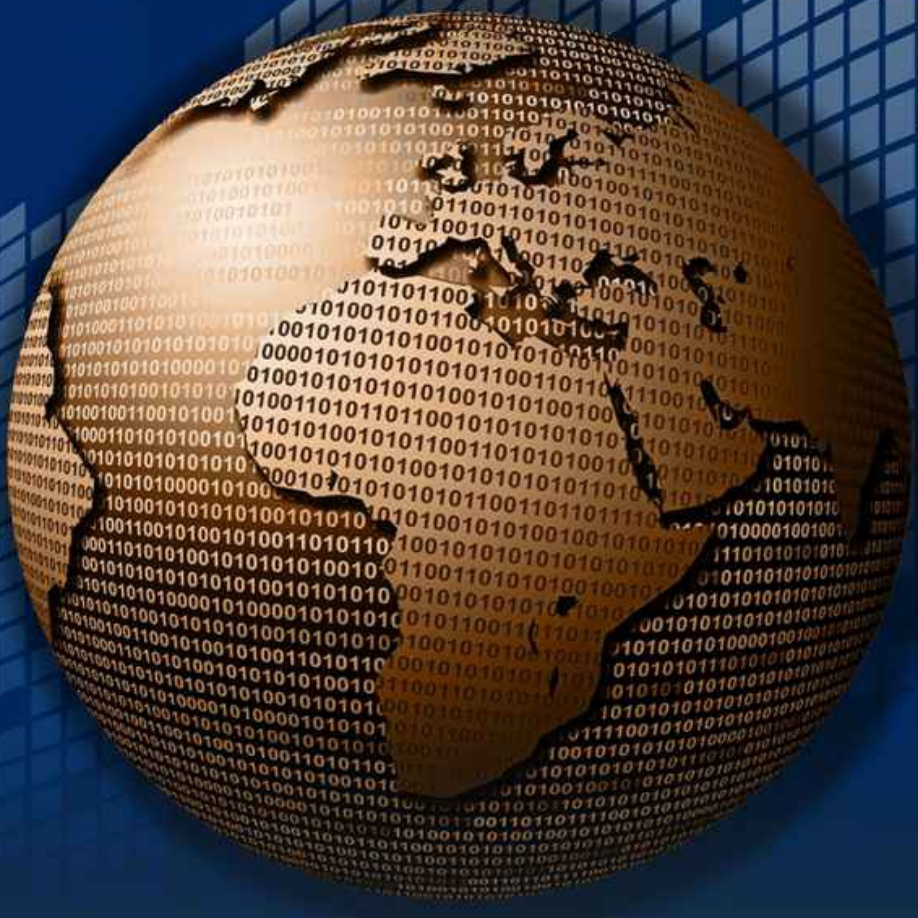
Cezar Taurion

velocidade | volume | variedade | veracidade | valor

BIG DATA

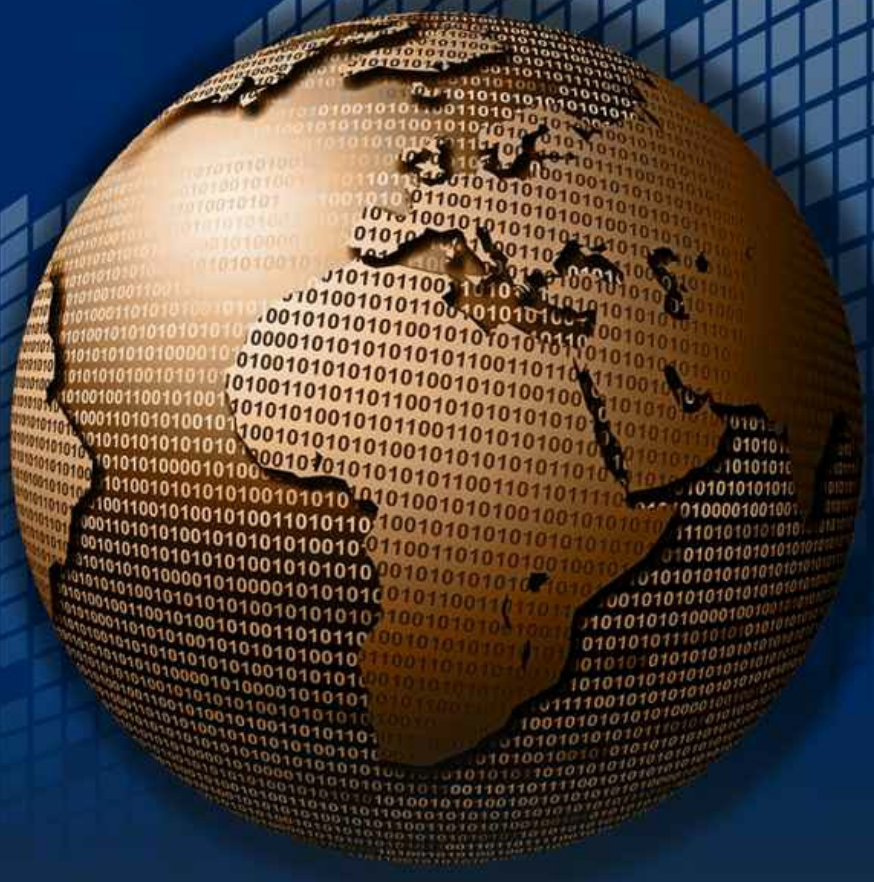


BRASPORT

Cezar Taurion

velocidade | volume | variedade | veracidade | valor

BIG DATA



BRASPORT

Para uma melhor visualização deste e-book sugerimos que mantenha seu software constantemente atualizado.

Editor: Sergio Martins de Oliveira

Diretora Editorial: Rosa Maria Oliveira de Queiroz

Gerente de Produção: Marina dos Anjos Martins de Oliveira

Assistente de Produção: Camila Britto da Silva

Revisão de Texto: Maria Helena A M Oliveira

Capa: usedesign

Produçao de e-pub: SBNigri Artes e Textos Ltda.

Taurion, Cezar

Big Data / Cezar Taurion - Rio de Janeiro: Brasport, 2013.

ePUB

T227b

ISBN: 978-85-7452-608-9

1. Big Data 2. Negócios I. Título.

CDD: 658.4

Ficha Catalográfica elaborada por bibliotecário – CRB7 6355

BRASPORT Livros e Multimídia Ltda.

Rua Pardal Mallet, 23 – Tijuca

20270-280 Rio de Janeiro-RJ

Tels. Fax: (21) 2568.1415/2568.1507

e-mails:

marketing@brasport.com.br

vendas@brasport.com.br

editorial@brasport.com.br

site: www.brasport.com.br

Filial

Av. Paulista, 807 – conj. 915

01311-100 – São Paulo-SP

Tel. Fax (11): 3287.1752

e-mail: filialsp@brasport.com.br

Sobre o Autor

Gerente de Novas Tecnologias Aplicadas/Technical Evangelist da IBM Brasil, é um profissional e estudioso de Tecnologia da Informação desde fins da década de 70. Com educação formal diversificada, em Economia, mestrado em Ciência da Computação e MBA em Marketing de Serviços, e consultor com experiência profissional moldada pela passagem em empresas de porte mundial, Taurion tem participado ativamente de casos reais das mais diversas características e complexidades tanto no Brasil como no exterior, sempre buscando compreender e avaliar os impactos das inovações tecnológicas nas organizações e em seus processos de negócio. Atuou em projetos de porte significativos, tendo inclusive sido líder da prática de IT Strategy para uma grande consultoria.

Escreve constantemente sobre tecnologia da informação em sites e públicações especializadas como CIO Magazine, Computerwold Brasil, Mundo Java, iMasters e TI Especialistas, além de apresentar palestras em eventos e conferências de renome. É autor de vários livros editados pela Brasport que abordam assuntos como Open Source/Software Livre, Grid Computing, Software Embarcado e Cloud Computing. Taurion também mantém um dos blogs mais acessados da comunidade developerWorks (www.ibm.com/developerworks/blogs/page/ctaurion). Este blog foi, inclusive, o primeiro blog da developerWorks na América Latina. Foi professor do MBA em Gestão Estratégica da TI pela FGV-RJ e da cadeira de Empreendedorismo na Internet pelo NCE/UFRJ.

Prefácio

Dados, o novo combustível dos negócios

Ao longo de quase quatro décadas de envolvimento com tecnologia, busquei continuamente fórmulas de como as empresas podem monetizar seus ativos de informação. Presenciei muitos conceitos e inovações promissoras virarem fracasso após consumirem altos investimentos e gerarem grandes expectativas que não se concretizaram. Em alguns casos, tive que esperar muito mais tempo para colher os benefícios, talvez pagando o alto preço pelo empreendedorismo e por utilizar de forma prematura conceitos e arquiteturas ainda não consolidados no mercado. Neste contínuo aprendizado, como é o atual assunto Big Data, vejo que esta inovação pode transformar por completo a maneira como trabalhamos e pensamos no desenvolvimento de novos produtos e novos serviços.

Apesar das transformações ocorrerem todos os anos, identifico uma questão em comum relacionada ao fascínio da engenharia da informação e, em especial com a arquitetura e modelagem de dados. Essas técnicas sempre se focaram em como coletar, organizar, colecionar e usar de forma competitiva os inúmeros dados disponíveis, e de como gerar valor para os negócios. Quando digo valor, leia-se resultados concretos em termos de novos negócios, novos produtos através da antecipação das necessidades dos clientes. Toda essa inteligência e conhecimento sobre o comportamento das pessoas, seus dados transacionais ou não, eventos correlacionados e seus relacionamentos em diferentes momentos e dimensões nunca foi tão discutida como recentemente. O papel dos líderes e executivos de tecnologia talvez nunca tenha sido tão discutido como agora.

Tudo leva a crer que chegamos neste ponto de forma pragmática, muito embora não haja um consenso a cerca do tema Big Data e o que isso realmente significa, sua profundidade e influências. Em linhas gerais, entendo que Big Data, além de ser mais uma buzzword (ou chavão de marketing), excede a capacidade atual que temos para usar os ambientes tecnológicos de hardware e as ferramentas de software para capturar, administrar e processar, incluindo-se correlacionar com alguma rapidez às informações que estão a nossa volta, provenientes de diferentes fontes. Muitas podem ser as fontes: dados de web e mídias sociais, ainda mais volumosos quando utilizados através de dispositivos móveis em função dos expressivos níveis

de crescimento mundial tanto no número de usuários como de aplicações; dados transacionais de diferentes naturezas (consumo, financeiro, gerenciamento de risco, seguros, telecomunicações, saúde, entretenimento); dados de biometria para identificação e validação (reconhecimento de voz, íris, retina, face); dados gerados no dia-a-dia pelas pessoas através de emails, logs/blogs, relatórios, apresentações e documentos em geral; e, comunicação entre dispositivos (móveis e fixos) que monitoram processos, máquinas, localizam pessoas, liberam acessos, contabilizam estoques, etc.

Estas fontes de dados hoje nos permitem pensar em aplicações que antes seriam impossíveis de serem desenvolvidas, pelo alto custo e pela complexidade envolvida. Vivemos e viveremos ainda mais essa complexidade em analisar estes eventos de forma a utilizar essas "deduções" analíticas na construção de novas ofertas de valor em produtos e serviços para nossos clientes.

Para que tenhamos respostas claras para a questão acima, devemos estar atentos a fatores relevantes quando pensamos em processar e analisar esse "tsunami" de informações provenientes do Big Data. Qual a relevância do volume de informação? Como processar essas informações com a velocidade necessária, com que capacidade de processamento e com que latência? Qual a variedade de dados (textos, som, vídeo, estruturados) que pretendemos trabalhar? Qual a procedência e a veracidade desses dados? Se são sensitivos, qual seu "prazo de validade" ou nível de atualização?

A maneira que vejo em encontrarmos respostas a estes fatores, concentra-se na forma como organizaremos nossas ações corporativas e como poderemos transformar as iniciativas de Big Data realmente rentáveis para nossas organizações.

Primeiramente recomendo a estruturação de um "ecossistema" de dados nas organizações, unindo a estratégia de dados tradicional ao Big Data, de forma integrada. Expanda sua estratégia de dados e analíticos e não descarte suas ferramentas atuais de data warehouse e business intelligence. Identifique como evoluir sua arquitetura tecnológica atual, pois ela continua e continuará sendo útil por muito tempo.

O segundo ponto que gostaria de destacar é o capital intelectual envolvido e sua organização, que sempre faz toda a diferença em tempo, custo e qualidade do resultado. Identifique as novas competências e habilidades necessárias. Prepare sua equipe. O conhecimento tecnológico é sempre importante, porém priorize o conhecimento de negócios para planejar e projetar os produtos e serviços para as operações da empresa. Atualmente muito se discute sobre o "cientista de dados" e o CDO (Chief Data Officer) e tantas outras novas competências. A Boa Vista Serviços foi a primeira empresa no Brasil a definir essa posição no contexto corporativo, focar sua atividade, atribuir o desafio estratégico de refletir

que dados são importantes para empresa e como rentabiliza-los. A informação é o principal ativo no gerenciamento de risco e de crédito. Fomos também a primeira empresa a fazê-lo de forma independente da organização de tecnologia da informação, que possui finalidade distinta, mas complementar.

Qualidade de dados é o terceiro ponto. É fundamental estruturar metodologicamente o processo operacional e os infindáveis atributos relacionados a validade, volatilidade, suas regras de transformação, seu processo de uso dentro da organização e como as aplicações de negócio utilizam esses dados de forma a gerar fator de diferenciação.

Como quarto fator, enfatizo a questão relativa à ética na captura e uso das informações, adicionando ao fator privacidade, onde fronteiras em termos de Big Data talvez ainda precisem ser debatidas e clarificadas. Recentemente o *World Economic Forum* lançou um projeto chamado "Rethinking Personal Data", que identifica possibilidades para crescimento econômico e benefícios sociais. Como parte dessa iniciativa, o Fórum definiu dados pessoais com "um novo ativo econômico", que abre muitas oportunidades e uma ampla discussão. Enquanto esse debate evolui e talvez consigamos identificar caminhos, ressalto que as questões de ética, na minha opinião, devem considerar prioritariamente: (a) identidade do proprietário dos dados; (b) sua privacidade; (c) a propriedade dos dados; e (d) nossa reputação, isto é, como somos percebidos e avaliados pelos dados analisados.

O uso de Big Data já deixou sua marca na história e seus impactos e implicações na sociedade e no mundo dos negócios. As empresas que souberem aplicar Big Data corretamente terão uma vantagem competitiva perante aos seus concorrentes.

Cesar Taurion tem percorrido uma trajetória ainda mais intensa e profícua no território da Engenharia da Informação, com ampla e sólida formação acadêmica, além de vivência nas engrenagens das operações de informática de grandes corporações. Desde cedo procurou delimitar e compreender as causas profundas dos empecilhos que têm povoado a vida prática dos responsáveis pela tecnologia da informação – em tempos em que exércitos de programadores atacavam dificuldades no processamento de dados por meio de calculadoras gigantes que antecederam os circuitos impressos e os microprocessadores empilhados nos waffers que sustentam as operações das nossas redes neurais do presente. Mais que tudo isso, Taurion é pioneiro na compreensão dos desafios para bem gerir bancos de dados construídos para interagir com demandas de negócios, uma solução que ganhou as pranchetas de fornecedores de aplicativos por volta de 1975 e que abriu as portas para a revolução da arquitetura distribuída e do computador pessoal. Com essa bagagem e uma capacidade de expressão que o faz conferencista reputado, aqui neste trabalho de grande percuciência e atualidade, circula pelas dificuldades intrínsecas da Era Big Data. E o faz com bastante

conhecimento de causa.

Tenho certeza de que este livro irá ajudar neste longo caminho que temos pela frente.

Boa leitura.

Dorival Dourado

Presidente & CEO

Boa Vista Serviços

Sumário

Bibliografia

Introdução

Outro dia me dei conta de quão profunda é a transformação digital que estamos vivendo. Este insight surgiu quando me peguei vendo televisão, falando com um amigo no Skype, mantendo três ou quatro chats simultâneos no Facebook e usando o Google para algumas pesquisas que complementavam estas conversações. Tudo isso ao mesmo tempo em que buscava comprar uma TV LED. E eu não sou um nativo digital, que multitarefa muitas vezes mais.

Na verdade, hoje estamos conectados digitalmente desde que acordamos até a hora de dormir, absorvendo um volume muito grande de conteúdo e também gerando muito conteúdo. Este fenômeno acontece no nosso dia a dia, seja em casa ou no trabalho. Aliás, esta é outra transformação. Fica cada vez mais difícil separar o "em casa" de "no trabalho". A computação está se tornando tão ubiqua que fica praticamente impossível separar o mundo físico do digital. Nos anos 90 (e isso tem menos de vinte anos) apenas os setores digitalizáveis como a música e a mídia tornaram-se digitais. No início dos anos 2000 o mundo físico se aproximou mais da digitalização com o comércio eletrônico e o Internet Banking. Hoje estamos começando a ver claros sinais da hiperconectividade, com cloud computing, a revolução da mobilidade e a Internet das Coisas permeando nossa sociedade. Nossos hábitos como pessoas conectadas, tornam-se hábitos como consumidores (checamos preços e avaliações antes de qualquer compra) e tornam-se também hábitos como funcionários (por que sou impedido de me conectar com os meus amigos via Facebook no escritório?)

Hoje quase um em cada sete habitantes do planeta está no Facebook. As plataformas de mídias sociais estão potencializando as nossas conexões. O Facebook não é apenas brincadeira de adolescentes ociosos. As pesquisas mostram que 72% da geração "baby boomer" já está nas mídias sociais.

O volume de informações geradas pela sociedade é assustador. Por exemplo, o Twitter sozinho gera diariamente doze terabytes de tuítes. Os medidores inteligentes de energia, que começam a substituir os seus ancestrais analógicos, geram 350 bilhões de medições por ano. E apenas uma pequena parcela das casas no mundo já tem estes medidores. Em 2010, Eric Schmidt, então CEO do Google disse que a cada dois dias a sociedade já gerava tanta informação quanto gerou dos seus primórdios até 2003, ou seja, cinco exabytes.

Smartphones e tablets já são lugar comum e surgiram há muito pouco tempo. O iPhone apareceu em 2007. Em março, a appStore da Apple alcançou a marca dos 25 bilhões de downloads de apps e agora em setembro foi a vez do Google Play. A mobilidade elimina as barreiras de tempo e espaço. As pessoas e leiam-se os consumidores ou melhor, os clientes, estão conectados todo o tempo e as empresas têm a oportunidade de estar em contato com eles também todo o tempo. Uma pesquisa da IBM com trinta mil consumidores em treze países mostrou que entre 78% e 84% deles se baseavam nas mídias sociais quando pensando em comprar produtos; 45% pediam opinião de amigos e/ou parentes; e apenas 18% se baseavam nas informações dos produtores e varejistas para a tomada de decisão. A conexão com o cliente é mais do que uma comunicação unidirecional. Não é mais suficiente apenas conectar-se a ele, mas é necessário integrá-lo aos processos de negócio e trazê-lo para dentro de casa.

O que isto significa? Que as informações sobre os produtos estão se tornando mais importantes que os produtos em si. Portanto, praticamente todas as empresas têm que estar no negócio de gerar conteúdo. Se o conteúdo gerado para a tomada de decisão de compra de um determinado produto não for adequado, a tendência da compra ser direcionada a outro produto torna-se bem maior.

Subestimar o impacto das transformações digitais pode colocar em risco o negócio, por mais sólido que ele pareça. Um exemplo é a indústria fonográfica que subestimou o poder da digitalização, ignorando que as conexões de banda larga ampliavam sua capacidade, o surgimento do padrão MP3 e os avanços tecnológicos que aumentaram de forma significativa a capacidade computacional dos chips ao mesmo tempo que a tecnologia se desmaterializava, miniaturizando-se continuamente. O resultado foi que de 2003 a 2012 o total da receita das gravadoras caiu de doze para oito bilhões de dólares. Por outro lado, a indústria da música como um todo gerou muito mais dinheiro, mas este dinheiro foi deslocado para outros atores do ecossistema que não as gravadoras, como produtores de shows, fabricantes de tocadores MP3 e assim por diante.

O que fazer diante desta revolução digital? As empresas podem ficar paradas esperando que as transformações passem por cima delas ou identificar oportunidades de aumentar seu espaço no mercado. Para isso devem agir mais rápido do que as concorrentes. Com o mundo hiperconectado podem redesenhar sua proposição de valor para seus clientes. Podem criar novos modelos de negócio onde produtos digitais substituam produtos físicos. Um exemplo é a mídia impressa que de ator principal passa a ser uma extensão da mídia digital, acessada por tablets. Os jornais terão de investir mais e mais em inovações no mundo digital. O resultante será uma mudança nas estratégias de negócio. O conteúdo digital, antes gratuito, está cada vez mais atrás dos chamados “paywalls”, ou seja, são pagos pelos seus assinantes.

Outras indústrias podem usar a tecnologia para otimizar seus produtos físicos. Por exemplo, a indústria automobilística usando sensores para detectar pontos cegos ao motorista. Esta mesma indústria pode criar novos serviços baseados no mundo digital, ofertando auxílio no trânsito, diagnósticos em tempo real e apoio em emergências.

A revolução digital também pode transformar a maneira como a empresa opera. Um exemplo interessante é o da Tesco, que criou um app móvel que permite aos clientes escanearem o código de barra dos produtos que querem comprar, seja em suas próprias casas ou que viram na casa de um amigo, e adicioná-lo automaticamente à sua lista de compras.

Ok, e como ir em frente? Minha sugestão: o primeiro passo é entender e identificar as oportunidades que a revolução digital pode trazer aos seus produtos e serviços. Não reagir contra. Depois, redesenhar a proposição de valor aos seus clientes. É a resposta ao "que fazer". Simultaneamente, construir o modelo operacional que sustentará esta nova proposição de valor. É o "como fazer". E executar e avaliar o processo continuamente. Em 2008, a indústria da mídia não imaginava que a ameaça dos tablets surgiria no ano seguinte. Monitorar as potenciais tecnologias disruptivas é essencial para a sobrevivência corporativa. Enfim, o desafio está à nossa porta. Temos a opção de pensar que a revolução digital não vai nos alcançar, o que com certeza será uma decisão errada, ou aproveitar a oportunidade!

Claro que fazer previsões é sempre um risco elevadíssimo. De maneira geral, a previsões falham porque não conseguimos identificar as informações realmente relevantes em meio ao ruído de dados e informações que nos cerca. Muitas vezes, limitados pelas nossas experiências, presumimos que a realidade atual vai se repetir indefinidamente. E não consideramos disrupções e quebra de paradigmas. Em fins do século XIX, o jornal londrino The Times previu que a sujeira dos cavalos soterraria Londres em menos de quarenta anos. Mas poucos anos depois surgiu o automóvel, que foi uma disrupção nos meios de transporte. Na área de TI, os últimos dez anos trouxeram muito mais mudanças que os cinquenta anos anteriores. Portanto, há dez anos atrás nenhuma tendência incluiria smartphones, tablets, Facebooks e Twiters (leia-se mídias sociais) e cloud computing. E, com a aceleração crescente das mudanças tecnológicas, as chances de acerto de qualquer previsão diminuem drasticamente!

Creio que nos próximos cinco anos ficará claro que a convergência tecnológica de quatro forças ou ondas que ainda estão em formação ou ainda mesmo são tsunamis em alto mar, estarão sobre nós, causando disrupções significativas na indústria de TI e no uso da tecnologia. Sim, falamos de cloud computing, mobilidade, social business e Big Data. Olhá-las de forma isolada é enganoso. Mas juntas provocam uma transformação na tradicional TI como nós conhecemos. A magnitude e a velocidade das mudanças é muito maior do que já vivenciamos em qualquer outra época da história da computação.

Em poucos anos passamos, por exemplo em cloud computing, de fazer perguntas do tipo "se vamos" para "quando e com que velocidade vamos". O debate agora é quando o modelo de computação em nuvem vai ultrapassar em volume de negócios o modelo on-premise atual. Nas plataformas sociais já temos mais de um bilhão e meio de pessoas e um bilhão se inseriu após 2009. Nesta velocidade é fácil imaginar um mundo onde os consumidores compram muito mais por dispositivos móveis do que por desktops, onde análises avançadas de dados, inclusive análises preditivas, tornam a maioria das decisões baseadas em fatos e em tempo real, onde cloud computing é o modelo computacional dominante (e deixaremos de falar cloud, para falarmos apena computing) e mais ideias e inovações surgirão das redes sociais que dos centros de P&D. Portanto, tecnologia é altamente estratégico para toda e qualquer organização.

Interessante que observo que ainda existe muita relutância em adotar estas tecnologias. Encontro algumas explicações para o fato. Uma é que os avanços tecnológicos têm se tornado tão rápidos que ultrapassam nossa capacidade de entendê-los e utilizá-los de forma diferente das que usamos hoje. Não reconhecemos a quebra de paradigmas que eles embutem. Thomas Kuhn no seu fantástico livro *"The Structure of Scientific Revolutions"* disse: *"Think of a Paradigm Shift as a change from one way of thinking to another. It's a revolution, a transformation, a sort of metamorphosis. It just does not happen, but rather it is driven by agents of change"*. Mas é difícil de perceber estas mudanças quando estamos no meio delas. Mais difícil ainda é começar a pensar de forma diferente quando todos os outros pares pensam sob o paradigma dominante. O efeito multidão é altamente inibidor. Apenas reconhecemos que o que temos não nos atende mais, mas ainda não percebemos que um novo paradigma já está sobre nós.

Outra explicação é a tradicional relutância diante do novo. Douglas Adams, famoso escritor de ficção, que foi o autor do conhecido "O Mochileiro das Galáxias" escreveu: *"Everything that's already in the world when you're born is just normal. Anything that gets invented between then and before you turn 35 is incredibly exciting and creative and you can probably make a career out of it. Anything that gets invented after you're 35 is against the natural order of things and the beginning of the end of civilization as we know it, until it's around for about 10 years, when it gradually turns out to be alright."*.

As áreas de TI que antes eram a porta de entrada da tecnologias nas empresas estão sendo sobrepujadas pelos usuários. Vem deles a adoção de tecnologias inovadoras e a força da chamada "consumerização de TI" é muito mais impactante do que parece à primeira vista. Na verdade desloca o eixo gravitacional da adoção de TI para fora da TI, pela primeira vez na história da TI corporativa. Esta nova geração de TI pode ser definida de forma simplista como de uso fácil e intuitivo, altamente móvel e social. Bem diferente da TI do teclado e mouse, que precisa esperar meses pela aquisição e entrada em operação de servidores físicos e armazena e trata as informações basicamente para atender aos sistemas transacionais. TI é hoje uma organização centralizadora, gerenciada por processos, que pastoreia seus usuários, definindo o que pode e o que não pode ser usado. Mas em cinco anos continuará assim? Pesquisas mostram que em 2016, 80% dos investimentos de TI envolverão diretamente os executivos das linhas de

negócio, e que estes serão os decisores em mais da metade destes investimentos.

Uma TI tradicional com seu imenso backlog de aplicações conseguirá justificar durante muito tempo todo este aparato quando, com um simples clique de um botão virtual em um tablet, pudermos fazer download de uma aplicação intuitiva e fácil de usar (dispensa manuais), contratar serviços de um aplicativo SaaS ou mesmo disparar um processo de criação de uma aplicação inovadora, como pode ser feito por serviços como o TopCoder (http://www.topcoder.com/)?

Hoje vejo que existem duas percepções diferentes. A TI olha a vinda destas tecnologias sob sua ótica tradicional e as tenta colocar sob o paradigma de comando e controle pela qual o próprio departamento de TI foi construído. Por outro lado, os usuários não querem mais ser tutelados desta forma. E aí creio que neste ponto é que veremos as tendências se consolidando nos próximos anos. Este tsunami tecnológico nos obrigará a buscar uma convergência das visões e percepções tanto de TI quanto dos usuários. Os extremos tentarão encontrar o ponto de equilíbrio. Mas, uma consequência para mim é indiscutível: a TI não poderá mais se manter burocrática e quase ditatorial como hoje. Se se mantiver indiferente ou contrária a estes movimentos, o termo "shadow IT" que hoje denomina a TI que corre por fora do controle da área de TI, impulsionada pelos usuários passará a ser a denominação da própria TI.

E aqui entramos em um dos quatro pilares, que é conhecido como Big Data. Segundo o IDC, em 2013, no universo digital, o total de dados criados e replicados será de quatro ZB, quase 50% mais que o de 2012 e quatro vezes o que foi em 2010.

O termo Big Data começa a despertar muita atenção, mas ainda é um conceito mal-definido e menos compreendido ainda. Com uma rápida pesquisa ao Google identifiquei um crescimento exponencial no interesse sobre o tema, mas pelo menos uma dúzia ou mais de definições. Big Data significa coisas diferentes para pessoas diferentes. A extensa cobertura da mídia, não apenas de tecnologia mas também de negócios, contribui para gerar mais confusão, pois as reportagens e textos abordam Big Data sob diversas óticas. Os fornecedores de tecnologia ajudam a aumentar a confusão pois cada um propaga seu viés, muitas vezes focado em simplifcar o processo, levando o leitor a considerar Big Data apenas como uma solução empacotada que pode ser colocada em prática simplesmente adquirindo tecnologia do próprio fornecedor.

Já existem, é claro, diversos casos de sucesso, mas a maioria das empresas ainda não tem uma visão clara do que é Big Data, do seu potencial e como alavancar esta potencialidade. Como o próprio conceito de Big Data ainda está um pouco nebuloso, o resultado é que a maioria das empresas ainda não percebeu o tsunami que é Big Data, porque ele ainda está em

alto mar. Mas rapidamente estará no litoral, provocando disrupções.

A observação do mercado nos mostra, portanto, que as empresas ainda estão nos estágios iniciais das iniciativas de Big Data, buscando compreender os conceitos e tecnologias por trás, para então começar a desenhar um road map para sua utilização.

Daí surgiu a ideia de escrever este livro, com o objetivo de não ser um profundo estudo sobre o tema, mas tentando apenas atender a este momento dos profissionais de TI e de negócios que estão em busca de uma melhor compreensão do que é Big Data e qual seu impacto nos negócios. A sua proposta é separar os fatos da ficção, ou seja, da realidade do hype que atualmente ronda o termo Big Data.

No livro vou abordar o assunto Big Data tanto do ponto de vista de negócios quanto das tecnologias envolvidas e debater alguns desafios que temos pela frente para colocarmos projetos de Big Data em ação. Antes de mais nada é importante lembrar que Big Data não trata apenas da dimensão volume, como parece à primeira vista, mas existe também uma variedade imensa de dados, não estruturados, dentro e fora das empresas (coletados das mídias sociais, por exemplo), que precisam ser validados (terem veracidade para serem usados) e tratados em velocidade adequada para terem valor para o negócio. A fórmula é então, Big Data = volume + variedade + velocidade + veracidade, gerando valor.

Volume chama atenção, mas é uma variável bem subjetiva, pois os limites de armazenamento e tratamento de dados aumentam com a rápida evolução tecnológica. Sem dúvida, quando falamos em volume hoje, os números já são gigantescos. Se olharmos globalmente estamos falando em zetabytes ou 10^{21} bytes. Grandes corporações armazenam múltiplos petabytes e mesmo pequenas e médias empresas trabalham com dezenas de terabytes de dados. Este volume de dados tende a crescer geométricamente e, em um mundo cada vez mais competitivo e rápido, as empresas precisam tomar decisões baseadas não apenas em palpites, mas em dados concretos. Assim, para um setor de marketing faz todo sentido ter uma visão 360° de um cliente, olhando não apenas o que ele comprou da empresa, como registrado no ERP, mas o que ele pensa e diz sobre a empresa, como o faz pelo Facebook e Twitter.

Portanto, os imensos amontoados de dados provêm das mais diversas fontes, pois, além dos dados gerados pelos sistemas transacionais das empresas, temos a imensidão de dados gerados pelos objetos na Internet das Coisas, como sensores e câmeras, e os gerados nas mídias sociais via PCs, smartphones e tablets. Integram o chamado Big Data o conteúdo de 640 milhões de sites, dados de seis bilhões de celulares e os três bilhões de comentários feitos diariamente no Facebook. Variedade porque estamos tratando tanto de dados textuais estruturados quanto não estruturados como fotos, vídeos, e-mails e tuítes. E velocidade, porque muitas vezes

precisamos responder aos eventos quase que em tempo real. Ou seja, estamos falando de criação e tratamento de dados em volumes massivos.

A questão do valor é importante. Big Data só faz sentido se o valor da análise dos dados compensar o custo de sua coleta, armazenamento e processamento. Existem também questões legais a serem resolvidas. Conheço um caso muito curioso de uma grande rede varejista americana que usa um sofisticado algoritmo de análise preditiva baseado na varredura de um imenso volume de dados de seus clientes. O algoritmo chegou à conclusão de que determinado padrão de compras e comentários nas mídias sociais levantava uma boa possibilidade de uma determinada pessoa estar grávida e enviou correspondência com promoções para grávidas para sua residência. Quem abriu foi o pai da adolescente que descobriu então a gravidez da filha. O advertising baseado nestas análises é uma questão ainda indefinida de invasão de privacidade. O uso de dados para prever eventos futuros da vida de uma pessoa tem consequências impactantes, particularmente se familiares ou potenciais empregadores passam a ter conhecimento de questões pessoais ligadas a estilo de vida ou estado clínico. Pior se a análise não for verídica, ocasionando um inconveniente muito grande e eventualmente um processo legal.

Outro desafio: criar e tratar apenas de dados históricos, com os veteranos Data Warehouse e as tecnologias de BI (Business Intelligence) começa a se mostrar um processo lento demais para a velocidade com que os negócios precisam tomar decisões. Aliás, o termo BI ou Business Intelligence já fez mais de cinquenta anos. Foi cunhado por Hans Peter Luhn, pesquisador da IBM em um artigo escrito nos idos de 1958 (http://en.wikipedia.org/wiki/Hans_Peter_Luhn).

Hoje já é consenso que dados são os recursos naturais da nova revolução industrial. Na atual sociedade industrial ter apenas recursos naturais como minério e exportá-los de forma bruta, importando em troca produtos manufaturados com eles não garante a competitividade de um país no longo prazo. O importante é a tecnologia e o conhecimento que cria produtos manufaturados. Afinal um quilo de satélite vale imensamente mais que um quilo de minério de ferro.

Fazendo um paralelo, na sociedade da informação é crucial saber tratar os dados na velocidade adequada. Dados não tratados e analisados em tempo hábil são dados inúteis, pois não geram informação. Na prática estima-se que cerca de 90% dos dados digitais disponíveis não estejam sendo adequadamente aproveitados. Dados passam a ser ativos corporativos importantes e como tal podem e deverão ser quantificados economicamente.

Esta importância se reflete no fato de que a primeira parte do livro aborda exatamamente as mudanças no cenário de negócios, bem como a criação de novos modelos de negócio baseados

no Big Data.

Big Data representa um desafio tecnológico pois demanda atenção à infraestrutura e tecnologias analíticas. O processamento de massivos volumes de dados pode ser facilitado pelo modelo de computação em nuvem, desde, é claro, que este imenso volume não seja transmitido repetidamente via Internet. Só para lembrar, os modelos de cobrança pelo uso de nuvens públicas tendem a gerar processamentos muito baratos mas tornam caras massivas transmissões de dados.

Indiscutivelmente que Big Data é um grande desafio para os CIOs e a área de TI. Primeiro temos as tecnologias que envolvem Big Data. Muitas vezes será necessário ir além das tecnologias tradicionais de banco de dados e Data Warehouse, entrando no campo dos bancos de dados NoSQL e processamento massivo. A principal base tecnológica para Big Data Analytics é o Hadoop e os bancos de dados NoSQL, onde No significa Not Only SQL, ou seja, usam-se bases de dados SQL e não SQL. A importância do "Not Only" SQL explica-se pelo modelo relacional ser baseado no fato de que, na época de sua criação, início dos anos 70, acessar, categorizar e normalizar dados era bem mais fácil que hoje. Praticamente não existiam dados não estruturados circulando pelos computadores da época. Também não foi desenhado para escala massiva nem processamento extremamente rápido. Seu objetivo básico era possibilitar a criação de queries que acessassem bases de dados corporativas e portanto estruturadas. Para soluções Big Data tornam-se necessárias várias tecnologias, desde bancos de dados SQL a softwares que utilizem outros modelos, que lidem melhor com documentos, grafos, processamento paralelo etc.

Mas este será um dos temas da segunda parte do livro, onde iremos explorar de forma mais abrangente as tecnologias envolvidas, inclusive discutindo mais detalhadamente um conceito muito interessante que é o stream processing. A ideia de stream processing ou stream computing é fantástica. É um novo paradigma. No modelo de data mining tradicional, uma empresa filtra dados dos seus vários sistemas e, após criar um data warehouse, dispara "queries". Na prática, faz-se garimpagem em cima de dados estáticos, que não refletem o momento, mas sim o contexto de horas, dias ou mesmo semanas atrás. Com stream computing esta garimpagem é efetuada em tempo real. Em vez de disparar queries em cima de uma base de dados estática, coloca-se uma corrente contínua de dados (streaming data) atravessando um conjunto de queries. Podemos pensar em inúmeras aplicações, sejam estas em finanças, saúde e mesmo manufatura. Vamos ver este último exemplo: um projeto em desenvolvimento com uma empresa de fabricação de semicondutores monitora em tempo real o processo de detecção e classificação de falhas. Com stream computing as falhas nos chips sendo fabricados são detectadas em minutos e não horas ou mesmo semanas. Os wafers defeituosos podem ser reprocessados e, mais importante ainda, podem-se fazer ajustes em tempo real nos próprios

processos de fabricação.

Temos também a questão da privacidade e acesso a dados confidenciais. É essencial criar uma política de acesso e divulgação das informações. Análises preditivas podem gerar resultados que podem questionar algumas perspectivas de negócios da empresa e precisam ter sua divulgação filtrada para não serem disseminadas inadequadamente. A capacidade analítica para traduzir dados em informações e conhecimento é outro desafio. Requer capacitação e ferramentas de visualização bem mais sofisticadas. Recomendo acessar o ManyEyes da IBM (http://www.958.ibm.com/software/data/cognos/manyeyes/) para uma ideia dos experimentos em novas formas de visualização de dados. Outra mudança que Big Data embute é a transformação das relações entre TI e o negócio. TI deve prover a base tecnológica, governança e procedimentos de segurança para o Big Data, mas as consultas e análises deverão ser feitas pelas áreas de negócio. O modelo Big Data deve ser, para o usuário, um modelo intuitivo de acesso, obtido pelo modelo simples de autosserviço.

A complexidade do Big Data vem à tona quando lembramos que não estamos falando apenas de armazenamento e tratamento analítico de massivos volumes de dados, mas de revisão ou criação de processos que garantam a qualidade destes dados e de processos de negócio que usufruam dos resultados obtidos. Portanto, Big Data não é apenas um debate sobre tecnologias, mas principalmente como os negócios poderão usufruir da montanha de dados que está agora à sua disposição. Aí emerge a questão da integração: como integrar bases de dados estruturadas e não estruturadas, com diversos softwares envolvidos?

A terceira parte do livro aborda a questão da capacitação profissional. Big Data abre oportunidades profissionais bem amplas. A capacitação profissional é um fator importantíssmo. Recomendo a leitura do texto "7 new types of jobs created by Big Data" em http://www.smartplanet.com/blog/bulletin/7-new-types-of-jobs-created-by-big-data/682.
Mostra que existem várias possibilidades de capacitação no setor, tanto no viés técnico como no analítico. No lado analítico é necessário preparação para sair do questionamento atual, que se faz pelo tradicional BI ("qual foi nossa taxa de crescimento em vendas mês a mês, nos últimos dois anos"), obtida pelos dados históricos armazenados no Data Warehouse e coletados pelos sistemas transacionais para novos tipos de análises. Com dados coletados em tempo real, não apenas pelos sistemas transacionais, mas também de mídias sociais, bases de dados públicas e outras, inclusive que estão em sistemas internos, mas inaproveitados, podemos, por exemplo, chegar a respostas a "como podemos crescer 20% no ticket médio de nossos clientes nas lojas da Zona Sul do Rio de Janeiro, ao mesmo tempo em que a concorrência está inaugurando duas novas lojas?". Analisar o comportamento dos clientes e do mercado como um todo pode levar à identificação de oportunidades de lançamento de novos produtos, focalizados em determinados nichos deixados a descoberto pela concorrência.

Pelo viés dos negócios, um artigo interessante que foi publicado há poucos meses pelo Wall Street Journal, edição brasileira, aponta como problema a escassez de profissionais conhecidos como data scientists (cientista de dados). O artigo diz que muitas empresas americanas começaram a procurar profissionais que saibam interpretar os números usando a análise de dados, também conhecida como inteligência empresarial. Mas, encontrar profissionais qualificados tem se mostrado difícil. Daí que várias faculdades americanas, como a Faculdade de Pós-Graduação em Administração da Universidade Fordham e a Faculdade de Administração Kelley, da Universidade de Indiana, começam a oferecer disciplinas eletivas, cursos de extensão e mestrados em análise de dados. Já o perfil mais técnico, o Data Architect, deve lidar com tecnologias SQL e NoSQL, conhecer profundamente conceitos como stream processing e Event Driven Architecture (EDA) e portanto ter capacidade de desenhar estratégias para manusear e analisar massivos volumes de dados de formatos diferentes quase em tempo real. Na minha opinião existe espaço para estes diversos perfis profissionais, tanto os mais voltados a negócios, qualificados para tratar analiticamente as informações geradas por estas imensas bases de dados e quanto para os de viés mais técnico.

Big Data deve começar a aparecer na tela do radar dos CIOs em breve. Aliás, já aparece no canto da tela de um ou outro CIO, e provavelmente em alguns anos já estará sendo persistentemente um dos temas mais prioritários das tradicionais listas de "tecnologias do ano" feitas pelos analistas de indústria. Portanto, é bom estar atento à sua evolução e eventualmente começar a colocar em prática algumas provas de conceito.

Na minha opinião Big Data é um ponto de inflexão que embute tanto potencial de mudanças quanto a nanotecnologia e a computação quântica. Os desafios que Big Data ainda apresenta são inúmeros, mas o principal é a falta de expertise e skills para lidar com o conceito e suas tecnologias. A demanda por novas funções como Data Scientist começarão a exigir respostas rápidas da academia. Big Data demanda conhecimento em novas tecnologias e principalmente mudanças no mindset da empresa. Seu valor está diretamente relacionado com o conceito de "empresa aberta", ou seja, a empresa sem silos entre departamentos e mesmo aberta a conexões com clientes e parceiros.

Em resumo, se a área de TI não quiser ser relegada a um simples departamento operacional deverá ser redesenhada. Deve entender, adotar e aceitar o papel de liderar as transformações que a tecnologia está e estará exercendo sobre as empresas nos próximos anos. Portanto, a principal tendência para os próximos anos, na minha opinião, é a mudança do papel de TI, passando a ser o impulsionador das transformações de negócio e não mais um centro de custo subordinado ao CFO ou ao diretor administrativo. Big Data tem papel fundamental nesta transformação.

Big Data tem o potencial de transformar economias, criando uma nova onda de produtividade econômica. Quando chegarmos a um cenário de "faça você mesmo" ou "do-it-yourself analytics", aí sim, Big Data será bem mais disseminado e útil para a sociedade. Na verdade, praticamente todo ramo de conhecimento humano vai ser intensivo em dados. Imaginemos ciência política, por exemplo. Com análise de centenas de milhões de dados gerados por posts em blogs, buscas por determinados assuntos, tuítes e comentários no Facebook, aliados a informações oficiais como press releases e artigos da mídia podemos analisar tendências de disseminação de determinadas correntes políticas, antes mesmo que pesquisas como as feitas tradicionalmente pelos institutos de pesquisa as apontem. Com uso de ferramentas automatizadas e novas formas de visualização atuando em cima de volumes medidos em petabytes, provavelmente não será mais necessário fazerem-se pesquisas de campo como feitas hoje.

Enfim, creio que Big Data está hoje onde a Internet estava em 1995, ou seja, quando começou a onda da Web e as primeiras iniciativas de comércio eletrônico surgiram. Ninguém conseguia prever, naquela época, o nascimento de empresas bilionárias como uma Amazon (criada justamente em 1995), de um Google (surgiu em 1998) e muito menos de um Facebook (2004), bem como as grandes mudanças que a Web provocou na sociedade. Portanto, acredito que apenas em torno de 2020 teremos uma ideia bem mais precisa do que as novas oportunidades de compreensão do mundo geradas pelo Big Data já estarão provocando nas empresas e na própria sociedade. Mas, os primeiros passos devem ser dados agora, sabendo-se dos riscos, mas também dos grandes prêmios do pioneirismo para as empresas que começarem primeiro. Teremos tempos excitantes pela frente!

O livro, como vocês verão, não é muito extenso. Optei por apresentar os conceitos de Big Data de forma resumida, atendendo às mudanças nos próprios modelos de negócios das editoras, com o advento e disseminação dos e-books. Agora podemos escrever livros mais sucintos, nos concentrando apenas no cerne da questão, sem necessidade de floreios para esticar o número de páginas que os modelos anteriores demandavam. Inseri no texto muitos links que ajudarão o leitor a se aprofundar em algum tópico que tenha mais interesse. Estes links levam a artigos e blogs que descrevem os temas com mais profundidade e recomendo que sejam acessados, pois contêm informações complementares bem relevantes. O uso intenso de links complementares ajuda a diminuir o tamanho do livro e permite uma leitura mais rápida, possibilitando que o leitor se aprofunde apenas nos temas em que tiver mais interesse.

Lembro também que as opiniões expressas neste livro são fruto de estudos, análises e experiências pessoais, não devendo, em absoluto, serem consideradas como opiniões, visões e ideias de meu atual empregador, a IBM, nem de seus funcionários. Todas as informações aqui fornecidas foram obtidas de livros e artigos disponíveis livremente para acesso ou aquisição na

Internet. Em nenhum momento falo em nome da IBM, mas apenas e exclusivamente em meu nome.

PARTE I

Conhecendo Big Data

Capítulo 1

O que é Big Data

Vamos começar este livro sobre Big Data explicando o que é e o que não é Big Data. Curiosamente, quando abordamos o tema surgem comentários do tipo “mas Big Data não é apenas um grande data warehouse?” ou “Big Data não é apenas um BI em cima de um data set de terabytes de dados?”. Sim, ambas são corretas, mas Big Data é muito mais que isso.

É certo que, indiscutivelmente, estamos falando de um volume de dados muito significativo. Mas, além de volumes abissais, existem outras variáveis importantes que fazem a composição do Big Data, como a variedade de dados, uma vez que coletamos dados de diversas fontes, de sensores, a ERPs e comentários nas mídias sociais, e a velocidade, pois muitas vezes precisamos analisar e reagir em tempo real, como na gestão automatizada do trânsito de uma grande cidade. Estas variáveis mudam a maneira de se analisar dados de forma radical. Em tese, em vez de amostragens, podemos analisar todos os dados possíveis. Um exemplo? Em vez de uma pesquisa de boca de urna nas eleições, onde uma pequena parcela dos eleitores é consultada, imaginem consultar todos os eleitores. Em teoria, é praticamente quase que a própria eleição.

Pessoalmente adiciono outras duas variáveis que são: veracidade dos dados (os dados tem significado ou são sujeira?) e valor para o negócio. Outra questão que começa a ser debatida é a privacidade, tema bastante complexo e controverso.

Existem diversas definições. Por exemplo, a McKinsey Global Institute define Big Data como “A intensa utilização de redes sociais online, de dispositivos móveis para conexão à Internet, transações e conteúdos digitais e também o crescente uso de computação em nuvem tem gerado quantidades incalculáveis de dados. O termo Big Data refere-se a este conjunto de dados cujo crescimento é exponencial e cuja dimensão está além da habilidade das ferramentas típicas de capturar, gerenciar e analisar dados”.

O Gartner, por sua vez, define como Big Data o termo adotado pelo mercado para descrever problemas no gerenciamento e processamento de informações extremas as quais excedem a capacidade das tecnologias de informações tradicionais ao longo de uma ou várias dimensões. Big Data está focado principalmente em questões de volume de conjunto de dados extremamente grandes gerados a partir de práticas tecnológicas, tais como mídia social, tecnologias operacionais, acessos à Internet e fontes de informações distribuídas. Big Data é essencialmente uma prática que apresenta novas oportunidades de negócios.

Com a revolução digital estamos diante da possibilidade de analisar um volume inédito de dados digitais, o fenômeno chamado Big Data, que para as empresas provavelmente terá um impacto tão grande em seus processos de negócio e decisão quanto a popularização da Internet. Os dados obtidos pelos sistemas transacionais, os visíveis comumente, representam uma parcela ínfima dos dados que circulam pelas empresas. Este imenso volume de "shadow data" ou dados invisíveis, tem passado despercebido e não é utilizado para melhorar os processos ou tomada de decisões. Big Data pode ser visto como a descoberta do microscópio, que abriu uma nova janela para vermos coisas que já existiam, como bactérias e vírus, mas que não tínhamos conhecimento. O que o microscópio foi para a medicida e a sociedade, o Big Data também o será para as empresas e a própria sociedade.

As informações vêm de todos os cantos. Vêm dos mais de seiscentos milhões de web sites, vêm dos cem mil tuítes por minuto, dos compartilhamentos de mais de um bilhão de usuários do Facebook que geram pelo menos 2,7 bilhões de comentários diariamente, dos sensores e câmeras espalhados pelas cidades monitorando o trânsito e a segurança pública, do um bilhão de smartphones...

Em agosto de 2012, o Facebook divulgou ao blog TechCrunchque processa 2,5 bilhões de conteúdo e mais de quinhentos terabytes de dados por dia. Com 2,7 bilhões de "curtidas" (termo adaptado para o português que se refere ao botão "curtir", específico da rede) e trezentos milhões de fotos por dia. Isso garante uma taxa de 105 terabytes a cada meia hora, permitindo que se tenha uma noção da quantidade de informações armazenadas pela rede.

Estima-se que dos 1,8 zetabytes (10^{21} bytes) gerados em 2012 pularemos para 7,9 zetabytes em 2015. Na verdade, cerca de 90% dos dados que existem hoje foram criados nos últimos dois anos. Em resumo, estamos diante de uma verdadeira avalanche de informações. A própria disseminação de novos meios de divulgação de informações como fotos e vídeos contribui para o crescimento do volume de dados. Um único segundo de um vídeo em alta-definição ocupa duas mil vezes mais bytes que o requerido para armazenar uma página de texto. A velocidade com que o mundo torna-se digital é assustadora. Em 2000, apenas 25% dos dados do mundo estavam armazenados em formato digital e em 2007 já eram 94%. Hoje deve estar bem próximo

dos 99,9%. Não significa que não existam mais dados em formatos analógicos, como em papel, mas o volume de dados digitalizados cresce de forma assombrosa. Olhando, por exemplo, a medicina, para cada página de um laudo médico (que pode estar tanto em papel como muito provavelmente também em formato digital) pode haver uns duzentos megabytes de imagens obtidas de aparelhos de raios-x, ressonância magnética e outros dispositivos. A crescente miniaturização da tecnologia bem como o aumento da sua capacidade de processamento e armazenamento permite a criação da Internet das Coisas, o que aumentará de forma exponencial a geração de dados.

A Internet das Coisas vai criar uma rede de centenas de bilhões de objetos identificáveis e que poderão interoperar uns com os outros e com os data centers e suas nuvens computacionais. A Internet das Coisas vai aglutinar o mundo digital com o mundo físico, permitindo que os objetos façam parte dos sistemas de informação. Com a Internet das Coisas podemos adicionar inteligência à infraestrutura física que molda nossa sociedade. A Internet das Coisas, com seus objetos gerando dados a todo instante, é um impulsionador poderoso para Big Data. Uma turbina de um moderno avião comercial a jato gera cerca de um terabytes de dados por dia, que devem ser analisados para mantê-la o maior tempo possível em operação. A diferença entre manutenção preventiva e preditiva pode ser vista neste caso. Uma manutenção preventiva diria que a cada x mil horas de vôo a turbina deve ser retirada da aeronave e revisada. Uma manutenção preditiva analisa os dados de operação gerada pelos sensores da turbina e prevê quando a manutenção deverá ser efetuada, podendo ser antes (evitando riscos para a segurança do vôo) ou depois, eliminando uma parada desnecessária (e caríssima) da aeronave.

Com tecnologias cada vez mais miniaturizadas podemos colocar inteligência (leia-se software) nos limites mais externos das redes, permitindo que os processos de negócio sejam mais descentralizados, com decisões sendo tomadas localmente, melhorando o seu desempenho, escalabilidade e aumentando a rapidez das decisões. Por exemplo, sensores que equipam um automóvel enviam sinais em tempo real para um algoritmo sofisticado em um processador no próprio veículo, que pode tomar decisões que melhoram a segurança da sua condução, evitando colisões ou mau uso dos seus componentes. Outras informações podem ser repassadas a uma central que monitore o percurso, gerenciando a forma do usuário dirigir o veículo e retribuir esta forma de direção em descontos ou taxas adicionais de seguros. Podem enviar informações que mostram que o veículo está sendo furtado e portanto decisões como o bloqueio de sua condução e acionamento da força policial podem ser tomadas.

A Internet das Coisas implica em uma relação simbiótica entre o mundo físico e o mundo digital, com entidades físicas tendo também sua única identidade digital, podendo com esta comunicar-se e interagir com outras entidades do mundo virtual, sejam estes outros objetos ou pessoas. Estamos, portanto, falando de muitos e muitos petabytes de dados gerados por estes

objetos.

A imensa maioria destes dados não é tratada e analisada. Mas já vemos alguns casos bem interessantes. A maior varejista do mundo, Walmart analisa diariamente trezentos milhões de comentários feitos por clientes no Twitter e Facebook. E com a computação em nuvem não é necessário dispor de um imenso parque computacional dentro de casa. Pode-se fazer esta análise em nuvens públicas e pagar-se apenas pelo consumo de recursos da análise. Isto permite criar o paradigma de Big Data sem a necessidade de possuir Big Servers ou servidores de grande capacidade.

Dados são os recursos naturais da sociedade da informação, como o petróleo para a sociedade industrial. Tem valor apenas se tratados, analisados e usados para tomada de decisões. Duas pesquisas reforçam esta afirmativa. Uma do Gartner coloca Business Analytics (BA) como uma das Top 5 prioridades dos CIOs nos últimos cinco anos e outra, da IBM, em um recente CIO Study que mostrou que 83% dos CIOs olham BA como altamente prioritário.

Gerenciar um negócio apenas na base de planilhas é manter a empresa sob um gerenciamento primitivo. Uma gestão por Excel não atende à complexidade e à velocidade que as decisões em um mundo cada vez mais complexo exigem. Os usuários estão móveis, com poderosos computadores de bolso como smartphones e tablets, as informações estão sendo processadas em nuvem e todo este aparato pode ser desperdiçado se não houver tecnologias que permitam levar informações analisadas para a ponta.

Um exemplo interessante de quão séria é a tecnologia foi a aquisição pela Walmart de um startup chamado Kosmix em 2011 (http://dealbook.nytimes.com/2011/04/19/wal-mart-buys-social-media-site-kosmix/). A proposta desta tecnologia é detectar, pela localização dos celulares dos clientes, o número de pessoas em cada loja e com essa informação os estoques das unidades que estão com vendas mais baixas são enviados para as que estão vendendo mais. Sugiro inclusive acessar o Walmart Labs (http://www.walmartlabs.com/) para ter uma ideia do que grandes varejistas estão investindo em tecnologias que agregam business analytics e o fator social ao e-commerce. Outro exemplo é a startup brasileira IDXP, ganhadora do programa de empreendedorismo Smart Camp da IBM, em 2012, com uma solução voltada à análise em tempo real do comportamento do cliente na loja (http://www.idxp.com.br/).

Resumindo, o que é Big Data? Vamos simplificar com uma simples fórmula para conceitualizá-lo. Big Data = volume + variedade + velocidade + veracidade, tudo agregando + valor.

Volume está claro. Geramos petabytes de dados a cada dia. E estima-se que este volume

dobre a cada dezoito meses. Variedade também, pois estes dados vêm de sistemas estruturados (hoje já são minoria) e não estruturados (a imensa maioria), gerados por e-mails, mídias sociais (Facebook, Twitter, YouTube e outros), documentos eletrônicos, apresentações estilo Powerpoint, mensagens instantâneas, sensores, etiquetas RFID, câmeras de vídeo etc. A variedade é um parâmetro importante pois, com diversas fontes de dados aparentemente sem relações, podemos derivar informações extremamente importantes e fazer análises preditivas mais eficientes. Por exemplo, conectando dados meteorológicos com padrões de compra dos clientes podemos planejar que tipo de produtos deverão estar em destaque nas lojas quando for detectado que haverá um período de alguns dias de temperatura elevada, daqui a três dias. Ou conectar dados geográficos com detecção de fraudes.

Velocidade porque muitas vezes precisamos agir praticamente em tempo real sobre este imenso volume de dados, como em um controle automático de tráfego nas ruas. Veracidade porque precisamos ter certeza de que os dados fazem sentido e são autênticos. E valor porque é absolutamente necessário que a organização que implemente projetos de Big Data obtenha retorno destes investimentos. Um exemplo poderia ser a área de seguros, onde a análise de fraudes poderia ser imensamente melhorada, minimizando-se os riscos, utilizando-se, por exemplo, de análise de dados que estão fora das bases estruturadas das seguradoras, como os dados que estão circulando diariamente nas mídias sociais.

E o que não é Big Data? Big Data não é apenas um produto de software ou hardware, mas um conjunto de tecnologias, processos e práticas que permitem às empresas analisarem dados a que antes não tinham acesso e tomar decisões ou mesmo gerenciar atividades de forma muita mais eficiente.

Capítulo 2

Por que Big Data?

Em 2012 a Pew Internet publicou um paper interessante, que pode ser acessado na sua íntegra em http://pewinternet.org/~/media//Files/Reports/2012/PIP_Future_of_Internet_2012_Big_Data.pdf O paper resume um estudo com centenas de pesquisadores e especialistas sobre os impactos, positivos e/ou negativos que Big Data poderá ocasionar nas empresas, pessoas e sociedade nos próximos anos. O crescimento no volume e variedade de dados é imenso e a velocidade de geração de novos dados está se acelerando rapidamente. Dados já começam a ser parte tão importante da economia como trabalho e capital. Assim, sairemos de uma era onde capital e trabalho determinavam os valores econômicos, para uma outra onde o valor será a conjunção do capital, trabalho e dados. Talvez entremos em uma época chamada de datanomics...

O relatório da Pew foi baseado nas respostas a duas questões, uma com viés positivo do efeito do Big Data: "Thanks to many changes, including the building of the Internet of Things, human and machine analysis of large data sets will improve social, political and economic intelligence by 2020. The rise of what is known as Big Data will facilitate things like newscasting (real-time forecasting of events); the development of inferential software that assesses data patterns to project outcomes; and the creation of algorithms for advanced correlations that enable new understanding of the world. Overall, the rise of Big Data is a huge positive for society in nearly all aspects". A outra questão apresentava viés negativo: "Thanks to many changes, including the building of the Internet of Things, human and machine analysis of Big Data will cause more problems that it solves by 2020. The existence of huge data sets for analysis will engender false confidence in our predictive powers and will lead many to make significant and hurtful mistakes. Moreover, analysis of Big Data will be misused by powerful people and institutions with selfish agendas who manipulate findings to make the case for what

they want. And the advent of Big Data has a harmful impact because it serves the majority (at times inaccurately) while diminishing the minority and ignoring important outliers. Overall, the rise of Big Data is a big negative for society in nearly all respects".

O resultado foi de 53% favorável ao viés positivo e 39% concordando com os posicionamentos negativos.

Este resultado mostra que estamos ainda no início da curva de aprendizado do que é Big Data e de seu impacto na sociedade. Pessoalmente, opto pelo viés positivo, mas com fortes alertas a questões como privacidade e uso indevido e não autorizado de informações pessoais. Mas, olhando o viés positivo, já existem casos bem interessantes de uso de Big Data, onde identificam-se neste oceano de dados padrões de conexões e interdependências que não conseguíamos observar quando usando amostragens bem menores. Um deles é o Flu Trends do Google. Baseado na imensa quantidade de dados que obtém a cada minuto no seu buscador e que estão relacionados com as necessidades das pessoas, o Google desenvolveu um projeto onde extrapolando-se a tendência de buscas conseguiu-se identificar tendências de propagação de gripe antes dos números oficiais refletirem a situação. Este tipo de previsão também pode ser feito para inflação, taxa de desemprego etc.

Existem diversos exemplos reais de uso de Big Data como os que a Amazon e NetFlix fazem com seus sofisticados sistemas de recomendação. Indico a leitura de um interessante artigo que conta o case da Intuit, empresa americana de software financeiro para pessoas físicas e pequenas empresas, que inclusive criou uma vice-presidência intitulada Big Data, Social Design and Marketing. Instigante, não? Vejam o link http://www.forbes.com/sites/bruceupbin/2012/04/26/how-intuit-uses-big-data-for-the-little-guy/.

Claro que a tecnologia ainda tem muito que evoluir, principalmente no tocante a maiores facilidades de manuseio de dados não estruturados e novas formas de visualização de dados. Veremos também um impulso maior na evolução das técnicas de Inteligência Artificial, como ferramenta auxiliar para a análise deste imenso volume de dados. Uma "learning machine" aprende com dados e quanto mais dados, mais o algoritmo aprende. Cria-se, portanto, um círculo virtuoso. Big Data é um passo significativo em busca da computação cognitiva. O exemplo do Watson da IBM é emblemático desta tendência. Recomendo acessar a página da IBM que descreve o Watson e suas aplicações em http://www-03.ibm.com/innovation/us/watson/index.html.

Adicionalmente, Big Data vai demandar novas funções e habilidades. Mas quando chegarmos a um cenário de "do-it-yourself analytics" ou um analítico modelo faça você mesmo, aí sim Big Data será bem mais disseminado e útil para a sociedade. Na verdade, praticamente

todo ramo de conhecimento humano vai ser intensivo em dados. Imaginemos ciência política, por exemplo. Com análise de centenas de milhões de dados gerados por posts em blogs, buscas por determinados assuntos, tuítes e comentários no Facebook, aliados a informações oficiais como press releases e artigos da mídia podemos observar tendências de disseminação de determinadas correntes políticas, antes mesmo que pesquisas como as feitas tradicionalmente pelos institutos de pesquisa as apontem. Com uso de ferramentas automatizadas e novas formas de visualização atuando em cima de volumes medidos em petabytes, provavelmente não será mais necessário fazer-se pesquisas de campo como são feitas atualmente.

Portanto, Big Data não é teoria ou futurologia. Geramos um imenso volume de dados a cada dia e análises de padrões e correlações nesta massa de dados podem produzir informações valiosíssimas em todos os setores da sociedade humana, de governos buscando entender demandas da população a empresas buscando se posicionar mais competitivamente no mercado.

Mas, observo que, quando se fala em Big Data, aparece uma concentração da atenção em análise e modelagem dos dados, herança das nossas já antigas iniciativas de Business Intelligence (BI). Claro que analisar os dados é fundamental, mas há outras etapas que merecem ser entendidas, para termos uma melhor compreensão do que é Big Data e seus desafios.

O que vejo são muitas empresas entrando em iniciativas de Big Data sem uma estratégia bem-definida que as oriente. Big Data não é apenas comprar pacotes de tecnologia, mas uma nova maneira de explorar o imenso volume de dados que circula dentro e fora das empresas. Big Data embute transformações em processos de negócio, fontes de dados, infraestrutura de tecnologia, capacitações e mesmo mudanças organizacionais na empresa e em TI.

A primeira fase de um processo de Big Data é a coleta de dados. Volume e variedade são suas caraterísticas. Estamos falando de coletar dados de sistemas transacionais, de comentários que circulam nas mídias sociais, em sensores que medem o fluxo de veículos nas estradas, em câmeras de vigilância nas ruas e assim por diante. Cada negócio tem necessidade de coletar dados diferentes. Uma empresa de varejo, por exemplo, demanda coleta de dados sobre sua marca, produtos e reputação nos comentários extraídos das mídias sociais. Um banco, querendo fazer uma análise de riscos mais eficiente ou uma gestão antifraudes mais aperfeiçoada, precisa não apenas juntar dados das operações financeiras dos seus clientes, mas também analisar o que eles comentam nas mídias sociais e até mesmo imagens obtidas de seu comportamento diante de uma ATM. Bem, começamos a levantar aqui as questões de privacidade.

Mas, coletar dados é apenas a primeira etapa. Um trabalho de limpeza e formatação também é necessário. Imaginemos uma imagem de raio-X de um paciente. Será armazenada da forma crua como obtida ou deverá ser formatada para ser analisada mais adequadamente depois?

Além disso é importante validar os dados coletados. Erros e dados incompletos ou inconsistentes devem ser eliminados para não contaminar as futuras análises.

Aí entramos em outra etapa, que é a integração e agregação dos dados obtidos das mais diversas fontes. Os dados dos sensores do fluxo de tráfego devem ser integrados aos dados dos veículos que estão transitando e mesmo com as de seus proprietários. Dados de diferentes tipos e formatos precisam receber tratamento específico. É importante definir categorias de dados e seus critérios de validação e aceitação. Também os critérios de segurança variam de acordo com as fontes de dados.

Depois desta integração temos então a fase mais visível que é a analítica, com a análise e interpretação dos resultados. É um desafio e tanto, pois terabytes de dados já existem e estão armazenados. A questão é "que perguntas fazer" para chegarmos à identificação de padrões e correlações que podem gerar valor para o negócio? Consultas ou queries em cima de um data warehouse gerado por transações obtidas pelo ERP são relativamente bem estruturadas e dentro de um domínio de conhecimento bem restrito. Mas, quando se coletam dados de diversas fontes, criar estas queries requer muito mais conhecimento e elaboração por parte dos usuários. É aí que entra o data scientist, um profissional multidisciplinar, com skills em ciência da computação, matemática, estatística e, claro, conhecimentos do negócio onde está inserido. Esta fase também demanda investimentos em pesquisas de novas formas de visualização, que ajudem a melhor interpretar os dados. Gráficos e planilhas tradicionais não são mais suficientes. Um exemplo interessante é o projeto ManyEyes da IBM (http://www-958.ibm.com/software/data/cognos/manyeyes/).

Big Data demanda também grande capacidade computacional. Um ambiente de computação em nuvem é bastante propício para suportar esta demanda. Para analisar volumes muito grandes é necessário também o uso de paralelismo, com tecnologias como Hadoop e MapReduce que abordaremos na parte 2. Um exemplo prático de uso de Hadoop associado a análise de dados é o Big Insights da IBM (http://www-01.ibm.com/software/data/infosphere/biginsights/).

Um desafio que precisa ser bastante debatido é a questão da privacidade. Muitos setores de negócios são altamente regulados como saúde e financeiro, por exemplo. Claro que a possibilidade de integrar dados das mais diversas fontes sobre um determinado indivíduo ou empresa é sempre uma fonte de preocupações. Imaginem o que cada um de nós deixa de pegada digital. Deixamos nossa pegada digital a todo momento, seja usando o Internet Banking, comprando pela Internet, acessando um buscador, tuitando, comentando alguma coisa no Facebook, usando o smartphone, ativando serviços de localização. Aglutinar todas estas informações permite a uma empresa ou governo ter uma visão bem abrangente daquela pessoa e de seus hábitos e costumes. Onde esteve a cada dia e o que viu na Internet. Se tem alguma

doença ou se tem propensão a sofrer de uma. Esta questão nos leva a outro ponto extremamente importante: garantir a segurança deste imenso volume de dados.

Estamos definitivamente entrando na era do Big Data. Talvez não tenhamos nos conscientizado disso e nem mesmo parado para pensar no potencial que volumes abissais de dados podem nos mostrar, se devidamente analisados e interpretados. Por outro lado, existem muitos desafios, com novas tecnologias, novos processos e novas capacitações. Enfim, temos muita ação pela frente e tanto para profissionais como para estudantes de computação (olha aí um tema quente para um trabalho de conclusão de curso ou TCC). Abre-se um novo e desafiador campo de atuação.

Enfim, creio que Big Data está hoje onde a Internet estava em 1995, ou seja, quando começou a onda da Web e as primeiras iniciativas de comércio eletrônico surgiram. Ninguém conseguia prever, naquela época, o nascimento de empresas bilionárias como uma Amazon (criada justamente em 1995), de um Google (surgiu em 1998) e muito menos de um Facebook (2004), bem como as grandes mudanças que a Web provocou na sociedade. Portanto, acredito que apenas em torno de 2020 teremos uma ideia bem mais precisa do que as novas oportunidades de compreensão do mundo geradas pelo Big Data provocarão nas empresas e na própria sociedade. Mas, os primeiros passos devem ser dados agora, sabendo-se dos riscos, mas também dos grandes prêmios do pioneirismo para as empresas que começarem primeiro. Teremos tempos excitantes pela frente!

Capítulo 3

Impactos do uso de Big Data

Big Data ainda está no canto da tela do radar dos executivos, mas tem o potencial de ser um disruptor de competitividade entre empresas. Afinal se uma empresa puder obter insights aprofundados sobre seus clientes, o que eles desejam e mesmo opinam sobre a empresa e seus produtos tem condições de mudar o jogo. Big Data e Analytics permitem encontrar padrões e sentido em uma imensa e variada massa amorfa de dados gerados por sistemas transacionais, mídias sociais, sensores etc.

Portanto, Big Data cria valor para as empresas descobrindo padrões e relacionamentos entre dados que antes estavam perdidos não apenas em data warehouses internos, mas na própria Web, em tuítes, comentários no Facebook e mesmo vídeos no YouTube. Isto foi reconhecido pela McKinsey em seu relatório "Big Data: The Next Frontier for Innovation, Competition and Productivity", acessível em http://www.mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_ne

O uso de Big Data já começa a se mostrar como um fator diferenciador no cenário de negócios. Alguns casos citados no relatório da McKinsey mostram que algumas empresas conseguiram substanciais vantagens competitivas explorando de forma analítica e em tempo hábil um imenso volume de dados. Big Data implica em duas palavras-chave: uma é volume (são bancos de dados de grandes volumes) e a outra é velocidade (o manuseio e tratamento analítico tem que ser feito muito rapidamente, em alguns casos até mesmo em tempo real). Isto se dá pela abrangência de dados que podem ser manuseados. Um Data Warehouse tradicional acumula dados obtidos dos sistemas transacionais como os ERP. Estes sistemas registram as operações efetuadas pelas empresas, como uma venda. Mas não registram informações sobre transações que não ocorreram, mas que de algum modo podem estar refletidas nas discussões

sobre a empresa e seus produtos nas mídias sociais. Também a empresa pode registrar diversas informações com a digitalização das conversas mantidas pelos clientes com os call centers e pelas imagens registradas em vídeo, do movimento nas lojas. Estas informações, geralmente não estruturadas, estão disponíveis e o que o conceito de Big Data faz é integrá-las de forma a gerar um volume muito mais abrangente de informações, que permita à empresa tomar decisões cada vez mais baseadas em fatos e não apenas em amostragens e intuição.

Mas colocar Big Data em prática não é simples questão de instalar alguma nova tecnologia. As tecnologias impulsionadoras são fundamentais, mas é necessário também que a empresa adapte seus processos de negócio de modo a explorar os insights gerados. Um exemplo de uso de Big Data de forma inovadora é a Pistoia Alliance (http://www.pistoiaalliance.org/), associação de empresas da indústria de "life sciences" que permite em modo de coopetição compartilhar dados para acelerar os seus processos de P&D. A ideia básica é criar um pool de Data Warehouses e através de processos inovadores e obviamente novas tecnologias, compartilhar informações entre diversas empresas. Utilizando-se modelos computacionais como cloud computing (aqui podemos falar em nuvens híbridas) ganha-se em economia de escala, permitindo que grupos de empresas possam implementar estratégias de Big Data que sozinhas não teriam condições financeiras e tecnológicas. Para isso é necessário criar padrões de acesso e regras e políticas bem definidas de privacidade e segurança de acesso. Aliás, privacidade e segurança em Big Data merece, pela importância do tema, um post específico.

Outro pré-requisito essencial é dispor de expertise, com novas funções como as de data scientist. O data scientist é um profissional multidisciplinar, com skills em ciência da computação, matemática, estatística e, claro, conhecimentos do negócio onde está inserido. Quem exatamente é esta figura não está claro, mas uma boa discussão pode ser vista em http://www.quora.com/Career-Advice/How-do-I-become-a-data-scientist. Também vale a pena dar uma olhada nesta comunidade: http://www.datascientists.net/.

Podemos resumir os impactos do Big Data, considerando que ele permite:

Maior transparência. A simples disponibilização de muito mais dados, antes inacessíveis, possibilita que o setor público, por exemplo, cruze informações antes isoladas em silos departamentais, abrindo novas oportunidades de integração e melhoria da gestão das cidades e órgãos. O que há de comum nas soluções a serem desenhadas para tornarem as cidades mais inteligentes, com mais segurança e melhor mobilidade urbana? A resposta é o uso otimizado das informações e tecnologias. As cidades que obtêm sucesso usam intensamente o conceito de cidades inteligentes como Singapura com seu tráfego inteligente ou New York com seu sistema de suporte à Polícia ou Crime Information Warehouse que pode ser visto em (ftp://ftp.software.ibm.com/software/solutions/pdfs/ODB-0144-01F.pdf). Para tratar este

imenso volume de informações hoje disponíveis, muitas geradas em tempo real, é necessário criar um centro de operações para a cidade. Este centro pode ser específico para um setor ou domínio, como operações de trânsito ou mais amplo, cross-domain, apresentando uma visão holística das informações sobre a cidade.

Segmentação bem mais precisa da população, chegando ao nível do próprio indivíduo. Com Big Data as fontes de informação se ampliam consideravelmente. Além disso, podemos chegar ao indivíduo. Capturando dados de rastreamento na Internet podemos inclusive dar outro sentido à palavra "anonimato". A capacidade cada vez maior de associarmos a identidade da vida real das pessoas com seus hábitos de compra marca uma virada na área de privacidade, desfazendo a fronteira cada vez mais nebulosa entre público o privado. Por exemplo, até segredos de leitura, antes indevassáveis para as editoras e livrarias, podem ser revelados agora. A Amazon, por exemplo, com seu leitor eletrônico Kindle, consegue obter informações sobre os hábitos de cada leitor. Ao ligá-lo ele envia para Amazon, entre outras informações, o livro que está sendo lido, quantas páginas foram lidas, o tempo consumido nesta leitura e os parágrafos sublinhados. A Barnes and Nobles, outra empresa que comercializa leitores eletrônicos, descobriu que livros de não ficção são lidos de forma intermitente, e os romances de forma contínua. Leitores de livros policiais são mais rápidos na leitura que os de ficção literária. Estas informações, impensáveis no mundo do papel, agora estão disponíveis para as editoras e livrarias desenvolverem ações comerciais como influenciar novos consumidores destacando os parágrafos mais lidos pelos demais.

Maior potencial de análises preditivas. Diversos projetos foram desenvolvidos usando-se informações coletadas de mídias sociais como Twitter e Google. Recentemente, abordando este assunto, li dois papers que me chamaram atenção para a crescente importância do fenômeno Big Data. O primeiro é "Big Data, Big Impact: New Possibilities for International Development", publicado pelo World Economic Forum, que pode ser lido em http://www3.weforum.org/docs/WEF_TC_MFS_BigDataBigImpact_Briefing_2012.pdf. O documento mostra como analisando padrões em imensos volumes de dados pode-se prever desde a magnitude de uma epidemia a sinais de uma provável ocorrência de uma seca severa em uma região do planeta. O documento mostra alguns casos muito interessantes, inclusive o projeto da ONU, chamado Global Pulse (http://www.unglobalpulse.org/), que se propõe a utilizar as tecnologias e conceitos de Big Data para ajudar a melhorar as condições de vida das populações do planeta.

Outro documento é "Obama administration unveils Big Data Initiative: announces US$ 200 million in new R&D investments", que pode ser acessado em http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release.pdf. O governo americano, ciente da importância do Big Data anuncia investimentos em pelo menos seis agências como National Science Foundation, National Institutes of Health e outras, para

investimentos em pesquisas na área de Big Data. Um dos projetos é o "1000 Genomes" que disponibiliza duzentos terabytes de dados para pesquisadores, armazenados em uma nuvem pública, neste caso da Amazon. O pesquisador paga apenas pelos serviços computacionais da nuvem e não pelo direito de acesso aos dados sobre genética humana.

Substituindo/complementando decisões humanas com algoritmos automatizados. Algoritmos sofisticados, suportados por imensos volumes de dados permitem automatizar diversas funções, como gerenciamento de processos, de tráfego nas ruas e assim por diante. O trânsito é um exemplo. Um dos grandes desafios das cidades é a mobilidade (ou imobilidade) urbana, representada pelo trânsito caótico e sistema de transporte público ineficiente. Uma pesquisa recente feita pela IBM, em vinte cidades do mundo inteiro, mostrou que o trânsito é uma das principais preocupações dos seus habitantes. O tempo de viagem entre dois destinos dentro de muitas das cidades mais importantes do mundo é um dos grandes motivos de insatisfação e stress, gerando imensas perdas em produtividade e qualidade de vida.

As seis cidades pior classificadas encontram-se nos países BRIC ou em desenvolvimento, como Beijing, Cidade do México, Johannesburg, Moscou, New Delhi e São Paulo. O rápido crescimento econômico dos países em desenvolvimento não foi acompanhado pela evolução da infraestrutura urbana e o resultado é um trânsito caótico. Por exemplo, somente nos quatro primeiros meses de 2010, Beijing registrou 248.000 novos veículos, uma média de 62.000 por mês. Na Cidade do México são emplacados duzentos mil novos veículos por ano e em São Paulo emplacam-se mais de mil veículos por dia.

O grande desafio é que, à medida que a urbanização aumenta, a tendência é a situação piorar cada vez mais. A ONU estima que em 2050 mais de 70% da população do mundo estará vivendo nas cidades, sobrecarregando mais ainda suas infraestruturas urbanas.

Claramente vemos que as soluções tradicionais, como construir mais avenidas e viadutos, está chegando ao seu limite. Não há mais espaço e nem tempo para longas intervenções urbanas, de modo que temos que buscar novas e inovadoras soluções.

A crescente disseminação da tecnologia está permitindo instrumentar e conectar objetos, possibilitando a convergência entre o mundo digital e o mundo físico da infraestrutura urbana. Com uma infraestrutura instrumentada e conectada podemos pensar em um sistema de transporte mais inteligente, conhecido pela sigla em inglês de ITS (Intelligent Transport Systems).

O conceito do ITS baseia-se na aplicação de tecnologias inovadoras para coletar mais e melhores dados, analisá-los de forma mais rápida e inteligente, e conectá-los através de redes mais eficientes para ações e decisões mais ágeis e eficazes. Uma mobilidade urbana mais

eficiente é crucial para a competitividade econômica de uma cidade. Alguns estudos apontam que congestionamentos intensos custam entre 1% a 3% do PIB das cidades.

Embora as consequências do congestionamento sejam similares, as suas causas e soluções são diferentes entre as cidades do mundo. Não se pode aplicar de forma automática uma solução que tenha sido bem sucedida em uma cidade em uma outra. Por exemplo, em Amsterdam mais de 50% das viagens dos cidadãos são ou a pé ou de bicicleta. Já em Chicago, 90% da movimentação é por carros particulares. Nas cidades dos países desenvolvidos a infraestrutura já existe e é necessária sua modernização. Na Europa o transporte público é o principal meio de mobilidade, enquanto que em muitas cidades dos EUA o carro é o principal meio. Já nas cidades dos países em desenvolvimento a infraestrutura está em construção. Também muitas destas cidades estão crescendo de forma muito rápida e desordenada.

Mas, como chegar a um sistema inteligente de transporte? É um processo de evolução gradual, que passa pela próprio nível de maturidade dos modelos de governança e gestão de transporte das cidades. Big Data permite coletar e analisar estes dados e em tempo real controlar as vias de acesso. Indo além, como em Singapura, fazendo análise preditiva de congestionamentos, o sistema consegue alertar sobre futuros congestionamentos com até 90% de precisão.

Criar novos modelos de negócio. Big Data permite a criação de novos modelos de negócio baseados no valor das informações armazenadas e analisadas. Empresas de diversos setores passam a ter condições, através de análises preditivas, de evitar o desperdício das manutenções preventivas. Por exemplo, porque um automóvel tem que trocar de óleo a cada cinco mil quilômetros? Baseado em algoritmos e bases de dados, as manutenções passam a ser preditivas, para cada veículo, pois o seu uso e consequente desgaste varia de acordo com os hábitos de seu motorista. Pode surgir neste meio uma empresa que colete dados de centenas de milhões de veículos de todos os tipos e marcas e venda estas análises para os fabricantes e empresas de seguro. Uma empresa de telefonia tem um imenso volume de dados sobre o comportamento de seus assinantes, ou seja, sabe para quem ligam, quando ligam e assim por diante. Podem então criar uma unidade de negócios para vender estas análises. Uma empresa pode analisar os dados de seus centenas de milhões de usuários de smartphones, identificar novas necessidades e vendê-las aos fabricantes, para que as coloquem nas futuras versões dos seus aparelhos.

Portanto, Big Data não é teoria ou futurologia. Geramos um imenso volume de dados a cada dia e análises de padrões e correlações nesta massa de dados pode produzir informações valiosíssimas em todos os setores da sociedade humana, desde governos buscando entender demandas da população até empresas buscando se posicionar mais competitivamente no mercado.

Capítulo 4

Alguns exemplos bem-sucedidos do uso de Big Data

Todos os setores de negócios serão afetados por Big Data, em maior ou menor grau. As empresas que conseguirem usar Big Data antes das concorrentes e se mantiverem inovadoras em sua utilização terão vantagens competitivas sustentáveis. A diferença no seu uso não está na tecnologia, uma vez que ela estará disponível a todas, inclusive a pequenas e médias, via computação em nuvem, permitindo criar cenários de Big Data sem Big Servers. O diferencial estará na sofisticação e maturidade da gestão da empresa.

Vamos exemplificar com alguns setores, como o setor da saúde, que envolve diversos atores, incluindo pacientes, profissionais da saúde, hospitais, laboratórios farmacêuticos, empresas de seguro saúde, governo e assim por diante. A área de saúde guarda muitas informações geradas por estes diversos atores. A possibilidade de integrar todas estas informações abre novas e surpreendentes perspectivas de inovação. O uso crescente de informações geradas por sensores que monitoram o paciente remotamente, o uso de informações que os próprios pacientes disponibilizam nas mídias sociais, aliados aos sistemas administrativos, clínicos (equipamentos médicos computadorizados) e financeiros permite criar um contexto inteiramente diferente do setor. Os médicos poderão trabalhar com informações dos hábitos dos seus pacientes fora dos hospitais, no seu dia a dia. O conceito do uso de informações muda de análises de fatos ocorridos, como crescimento do número de pessoas infectadas por dengue nos últimos anos, para uma análise preditiva, que poderá apontar quantas pessoas e em que locais estarão sendo infectadas pela dengue nos próximos meses, permitindo a tomada de decisões e ações preventivas muito mais eficazes.

Por exemplo, na área de pesquisa e desenvolvimento (P&D), analisando-se de maneira mais efetiva os dados, podem-se desenvolver medicamentos, tratamentos médicos e programas governamentais para saúde de forma muito mais eficiente. Um exemplo: analisando-se um volume significativo de dados pode-se ter uma ideia mais aprimorada da eficácia de determinados remédios ou tratamentos. Onde obter estes dados? Nas próprias clínicas e hospitais há uma infinidade de dados, a maioria dos quais não é analisada e muitas vezes simplesmente descartada. Uma varredura e um tratamento analítico nestas centenas de milhares de registros médicos nos ajudará a identificar correlações entre os tratamentos aplicados aos pacientes e os seus resultados. Além disso, com um volume significativo de dados médicos, sofisticados algoritmos preditivos podem modelar de forma mais eficiente em quais medicamentos alocar os recursos de P&D. Também no desenvolvimento de medicamentos, Big Data pode trazer grandes benefícios, pois as pesquisas frequentemente baseiam-se em conjuntos pequenos de dados, muitas vezes coletados depois que os medicamentos são introduzidos no mercado. Hoje em dia é cada vez mais comum os pacientes se encarregarem dos cuidados com sua própria saúde, fazendo pesquisas on-line sobre suas doenças, ingressando em redes onde outros pacientes estão também conectados, trocando experiências e ajudando uns aos outros, crescendo assim o volume de dados que podem ser analisados, facilitando a monitoração da eficácia dos medicamentos. Um exemplo é o http://www.PatientsLikeMe.com onde os próprios pacientes podem compartilhar suas experiências e tratamentos médicos e o http://www.Sermo.com, onde médicos trocam experiências sobre casos reais de tratamento com que estão envolvidos.

Um exemplo interessante do uso de Big Data na pesquisa de medicamentos é o cruzamento de informações entre usuários de diversas drogas. Suponhamos que uma pessoa que sofra de pressão alta tome determinado medicamento. Ela é alertada de efeitos colaterais como palpitações. Um outro indivíduo está se medicando com antidepressivos e também é alertado de efeitos colaterais. Mas, se uma pessoa está tomando os dois remédios ao mesmo tempo? Qual o feito colateral resultante? É onde entra Big Data. Pesquisadores da Universidade de Stanford estão usando técnicas de mineração de dados para fazer estas correlações. Estas análises já mostraram dezenas de combinações de remédios que geram efeitos colaterais anteriormente desconhecidos. Com um volume bem grande de dados é possível conseguir-se identificar relacionamentos que passariam despercebidos em pequena escala.

Outro uso potencial é no controle de doenças, aumentando a vigilância sobre possíveis surtos como, por exemplo, descobrindo em que áreas elas foram detectadas e de forma preditiva alertar e preparar os hospitais e centros de emergência regionais para uma possível epidemia. Pesquisas demonstram que isso é factível já havendo casos bem interessantes de uso de Big Data neste contexto, onde identificam-se tendências em um oceano de dados padrões de conexões e interdependências que não conseguíamos observar quando usando amostragens bem

menores. Um deles é o Flu Trends do Google. Baseado na imensa quantidade de dados que obtém a cada minuto no seu buscador e que estão relacionados com as necessidades das pessoas, o Google desenvolveu um projeto onde, extrapolando-se a tendência de buscas, conseguiu-se identificar indícios de propagação de gripe antes dos números oficiais refletirem a situação. Este tipo de previsão também pode ser feito para inflação, taxa de desemprego etc. Um paper que descreve em maiores detalhes o Flu Trends pode ser lido em http://static.googleusercontent.com/external_content/untrusted_dlcp/research.google.com/es//arc influenza-epidemics.pdf. Uma outra experiência utilizou o Twitter para determinar antecipadamente uma epidemia de cólera no Haiti, depois do terremoto em 2010, pelo menos duas semanas antes que os órgãos oficiais a detectassem. Assim, também para as doenças contagiosas, pode-se ter um controle das áreas geográficas de risco de incidência da doença, atuando-se de forma preditiva.

Os pacientes também podem utilizar informações mais adequadas a respeito dos custos, qualidade e eficiência de seus tratamentos e compará-los com os resultados obtidos por outros pacientes. Com estes dados pode-se fazer uma análise da eficiência e desempenho das entidades envolvidas na saúde, como hospitais e clínicas, considerando-se o atendimento, tempode espera e eficiência dos tratamentos propostos. O resultado é uma maior transparência do setor como um todo, permitindo que ações de melhoria sejam realizadas com maior eficiência.

Adicionalmente, associando-se Big Data com um mundo cada vez mais instrumentado, podemos utilizar com muito mais intensidade sensores que monitorem pacientes remotamente, analisando continuamente seus dados clínicos e através de algoritmos preditivos alertando os médicos de qualquer anomalia, antes que elas aconteçam.

Big Data nos permite criar novos modelos de negócio na área de saúde, como agregadores de conteúdo, que podem coletar e vender informações cruzadas sobre o setor. São negócios baseados na coleta e análise de informações especializadas. Algo similar ao modelo das empresas de TV a cabo que coletam e entregam conteúdo (através de contratos com provedores de conteúdo, como os canais de filmes), estes novos negócios poderão ser os funis através dos quais as informações essenciais serão disponibilizadas ao mercado. O diferencial competitivo para estes negócios será a relevância de conteúdo agregado e disponibilizado.

Outro setor que terá muito a ganhar com uso de Big Data é o setor da administração pública. Os governos de todo o mundo estão sendo cada vez mais pressionados a se tornarem mais eficientes e menos onerosos para a sociedade. Além disso, muitos países, como o Brasil criaram leis que tornam disponíveis informações públicas, como a lei de Acesso à Informação, a Lei 12.527 promulgada em 18 de novembro de 2011. Por outro lado, a administração pública ainda tem a maioria de seus órgãos atuando de forma isolada, em silos informacionais, onde a

troca de informações entre os setores e destes com a sociedade é precária e desestruturada. Mas, existe um aspecto bastante positivo. Embora atuando de forma isolada, os órgãos públicos vêm há tempos digitalizando seus dados e documentos. Por exemplo, o Imposto de Renda e as eleições são inteiramente digitais. Não circula mais papel nestas atividades. O resultado é que existem volumes significativos de informações que podem ser analisadas, permitindo que a administração pública torne-se mais eficiente. Por exemplo, detectando anomalias nos sistemas como arrecadação de impostos ou pagamentos de benefício. Com o Big Data, algoritmos sofisticados podem rastrear dados provenientes de diversas fontes identificando erros e fraudes. Com base em associações e cruzamentos de dados, podem-se identificar situações anômalas como, por exemplo, um funcionário público acumulando cargos indevidamente ou um cidadão que obtém salário desemprego ao mesmo tempo em que apresenta atestado de acidente de trabalho. Big Data permite uma maior transparência do setor público, possibilitando que a sociedade exija mais e mais eficiência na administração pública.

O potencial do Big Data na administração pública só será alcançada com mudanças organizacionais, culturais e de processos. Os governos passam a ter a possibilidade de tomar decisões baseados em fatos e operar com muito mais eficiência. Entretanto, o primeiro passo não é tecnológico, mas a decisão política de quebrar silos entre departamentos e esferas diferentes de governo. A tecnologia é apenas impulsionadora, mas as grandes transformações deverão ser organizacionais e culturais. Um passo importante é a abertura de dados antes restritos a setores públicos, à toda a sociedade. Mas, apenas disponibilizar, sem ferramentas de acesso fácil e intuitivas que permitam aos cidadãos navegar por este oceano de dados e fazer suas análises é limitante. Uma iniciativa interessante é a da cidade de New York, com o projeto NYC Digital (http://www.nyc.gov/html/digital/html/apps/apps.shtml), onde são disponibilizados diversos apps para smartphones e tablets para os cidadãos obterem informações e participarem ativamente da gestão da cidade. O NYC Digital incentiva diversas ações de engajamento da sociedade com a administração pública realizando eventos como hackathons e competições para desenvolvimento de apps para equipamentos móveis, voltados para melhoria da cidade.

Claro que abrir e integrar dados antes fechados para a sociedade gera angústias e receios, principalmente devido às características específicas de determinados dados da administração pública. É necessário criar uma política de segurança da informação que classifique os dados de acordo com seu nível de privacidade. Existem dados que não podem ser disponibilizados publicamente e outros que podem, e às vezes não o são apenas por questões culturais. Os dados de órgãos públicos não podem ser classificados de forma única, mas analisados caso a caso para poderem ser disponibilizados. Outro desafio é que, de maneira geral, a maioria dos órgãos de governo utiliza tecnologias diferentes, havendo uma barreira natural de acesso aos dados, que é a incompatibilidade dos formatos das bases de dados. Temos também a questão da

descrição dos dados, sendo importante criar um mecanismo de metadados que torne compreensíveis os dados tornados públicos.

Quando falamos da busca pela maior eficiência da administração pública nos vem à mente as questões que diretamente nos afligem, como habitantes das cidades. Estamos cada vez mais urbanos. A urbanização traz inúmeros benefícios para o desenvolvimento econômico. As cidades são centros econômicos de inovação, cultura, conhecimento, novas ideias e suas aplicabilidades. Existe uma clara e positiva correlação entre o crescimento econômico e o grau de urbanização de um país. Embora nem todo país urbanizado seja desenvolvido, não há um único país desenvolvido que não esteja altamente urbanizado. Portanto, sem sombra de dúvidas, as cidades são polos de atração para talentos e capital humano. Mas, por outro lado, a urbanização acarreta imensos desafios sociais e econômicos. Nas cidades dos países emergentes, como o Brasil, o crescimento rápido da economia e da urbanização gera uma pressão muito forte na infraestrutura, provocando problemas de trânsito, quedas de energia, bolsões de pobreza, criminalidade e deficiências nos sistemas de ensino e saúde.

Uma volta pelo Brasil nos mostra que as suas grandes cidades apresentam uma infraestrutura que não dá conta do seu crescimento. Em maior ou menor grau, os problemas são praticamente os mesmos. A densidade populacional cresce e este crescimento é desordenado. É um crescimento orgânico com as cidades se espalhando em termos de população e área geográfica.

Hoje, em algumas cidades brasileiras já se fala no "apagão da mobilidade", com seu trânsito caótico e engarrafamentos crônicos afetando a qualidade de vida e roubando recursos da economia. Segundo a Fundação Dom Cabral, estima-se que somente em São Paulo, os gargalos urbanos roubem quatro bilhões reais a cada ano.

Tentar resolver os problemas da maneira que comumente estamos acostumados, ou seja, apenas pelo lado físico, abrindo mais ruas e avenidas, construindo mais escolas e colocando mais policiais na rua não será suficiente. Nem sempre haverá espaço para abrir novas avenidas e nem sempre será possível obter orçamentos que aumentem significativamente a força policial. Além disso, uma nova avenida pode simplesmente resultar em maior volume de tráfego, aumentando o problema e gerando mais poluição. Mas é indiscutível que algo precisa ser feito urgentemente e por que não começarmos a criar uma urbanização mais inteligente?

Precisamos resolver os dilemas econômicos, sociais e ambientais que nortearão as políticas públicas de forma inovadora, quebrando hábitos arraigados e gerando novos modelos de uso da infraestrutura urbana.

A tecnologia e entre estas, Big Data, tem um papel fundamental neste processo

“revolucionário”. Por exemplo, algumas soluções inovadoras para transporte e trânsito já vêm sendo colocadas em prática, com sucesso, em cidades como Estocolmo, Londres e Singapura. Em Estocolmo, um novo sistema inteligente de pedágio reduziu de maneira impressionante os congestionamentos de tráfego e as emissões de carbono. Em Londres, um sistema de gerenciamento de congestionamentos reduziu o volume de tráfego a níveis da década de 1980. Em Singapura, um sistema pode prever velocidades no tráfego com precisão de 90%. Com algumas melhorias, o sistema vai poder também prever, em vez de apenas monitorar, outras condições do trânsito

Mas, por que fazer isso? Como as cidades são pólos econômicos, indiscutivelmente que começarão a competir entre elas pela atração de mais negócios e fazer crescer sua economia. Para atrair talentos e negócios é imprescindível uma infraestrutura de qualidade, que possibilite uma mobilidade urbana segura e adequada, que ofereça serviços de saúde e educação de bom nível, e que crie opções de lazer. Em resumo, ofereça qualidade de vida. As cidades deverão ser gerenciadas como empresas, visando crescimento econômico, mas aliando este crescimento à sustentabilidade e à qualidade de vida. A atratividade baseada única e exclusivamente em isenção de impostos e doação de terrenos para indústrias está se esgotando rapidamente.

A reengenharia do modelo de urbanização passa por um bom planejamento a longo prazo, perfeitamente conectado às inovações tecnológicas. A infraestrutura urbana deve ser baseada na convergência dos mundos analógicos e físicos, com o mundo digital.

Portanto, para exercerem seu papel de “motores da economia” a maioria das cidades deve assumir atitudes proativas e holísticas de melhoria de suas propostas de qualidade de vida para seus cidadãos; redesenhar os modelos obsoletos de gestão e processos de governança que na maioria das vezes não se alinham mais com a complexa sociedade em que vivemos e reconhecer o papel fundamental que as tecnologias como Big Data podem assumir nos seus projetos de urbanização sustentável. Um exemplo interessante é a cidade de Nova York que, através de Big Data e análise preditiva, investiga potenciais casos de riscos de incêndio. Nos primeiros seis meses de uso destas técnicas o Corpo de Bombeiros da cidade identificou riscos reais em 75% dos casos indicados pela análise. As cidades são os locais onde a sociedade mais pode interagir com a administração pública e a tecnologia, entre elas Big Data, que oferece o grau de transparência para que esta interação ocorra. No Reino Unido, o data.gov.uk permite acesso a mais de 5.400 bancos de dados que disponibilizam estatísticas de criminalidade, gastos governamentais e até mesmo taxas de infecção hospitalar, por hospital. Baseadas nestas informações, centenas de apps já foram criadas, facilitando o acesso dos cidadãos a estes dados, potencializando seu engajamento e por sua vez, demandando maior eficiência do setor público.

Outro setor onde Big Data afeta a sociedade é na segurança pública. Com dados coletados de diversas fontes, que vão de câmeras nas ruas a comentários e posts publicados em mídias sociais, as agências de inteligência e de segurança pública podem detectar e se antecipar a atividades ilícitas, evitando que ocorram. Ao utilizar tecnologias de Big Data as agências podem descobrir tendências comportamentais criminosas, cruzando pedaços de informações que aparentemente não estão correlacionados. Este cruzamento de informações é essencial para o combate preditivo às ações ilícitas. Um exemplo de Big Data em segurança pública vem da cidade de Nova York. Muitas vezes, o que leva à solução de um crime é um detalhe insignificante, como um apelido, uma tatuagem ou um simples recibo de estacionamento. E são bilhões de detalhes como esses que estão armazenados no banco de dados do New York City Real Time Crime Center (RTCC). Nova York é hoje a grande cidade mais segura nos Estados Unidos, um exemplo de como as cidades estão se tornando mais inteligentes em matéria de segurança pública. Hoje o RTCC reúne mais de 120 milhões de queixas criminais na cidade, 31 milhões de registros de crimes nacionais e 33 bilhões de informações públicas. Métodos sofisticados de análise de dados e recursos de busca estabelecem conexões através de múltiplos bancos de dados. As informações podem ser visualizadas, em segundos, em uma tela de vídeo com a altura de dois andares: a foto de um suspeito aparece com detalhes, como tatuagens, delitos anteriores, endereços com mapas, muito rapidamente. Dados críticos são enviados instantaneamente aos policiais na cena do delito e o que antes levava dias, agora é feito em minutos, com um ganho indiscutível para a sociedade.

Big Data na defesa civil é outra oportunidade a ser explorada. Durante um desastre natural, como uma enchente (felizmente não temos terremotos no Brasil) dados gerados por GPS embutidos em smartphones e sensores e câmeras que analisam o fluxo de veículos podem contribuir para facilitar a evacuação de pessoas das áreas atingidas, bem como diminuir o tempo para o socorro chegar a estas áreas.

Setores como o financeiro, cartões de crédito e o de seguros podem usar Big Data como uma poderosa arma no combate às fraudes. Para combater fraudes, quanto mais informações que possam ser cruzadas e analisadas, melhor. Além disso, a velocidade com que a fraude é detectada ou até mesmo prevista diminui seu impacto na empresa. De maneira geral, o modelo tradicional de combate às fraudes baseia-se na identificação de pessoas ou clientes que preencham determinados critérios. O problema é que este modelo funciona no atacado, mas não consegue distinguir casos individuais. Um exemplo prático: em todas as minhas férias compro com antecedência passagens aéreas em empresas low-fare estrangeiras. Os sistemas de prevenção de fraude dos cartões de crédito bloqueiam as minhas compras porque, embora compradas via Internet, o local das empresas aéreas é no exterior e consequentemente fora da minha usual região geográfica de compra. Todo ano acontece a mesma coisa e sou obrigado a ligar para solicitar a liberação durante o período em que estou desenhando o meu roteiro de

férias e comprando os bilhetes aéreos. Claro, é positivo, pois aumenta minha sensação de segurança de que nenhuma outra pessoa está usando meu cartão indevidamente. Mas, em três anos sucessivos as minhas férias foram no mesmo período e as passagens foram também compradas nos mesmos meses. E os sistemas não detectaram este padrão.

Outro desafio é que muitas vezes as fraudes são identificadas depois que as transações ocorreram, quando então o prejuízo é assumido pela empresa de cartão de crédito ou banco. Alguns estudos apontam que as empresas utilizam menos de 20% das informações disponíveis, seja dentro delas ou em outras bases de dados e mídias sociais para modelar e detectar possíveis fraudes. A questão é como coletar e analisar estes outros 80%? E identificar fraude no momento em que a tentativa está ocorrendo? E melhor, se pudermos prever futuras tentativas de fraude?

O setor bancário tem muito a ganhar com uso intensivo de Big Data. Hoje os bancos trabalham basicamente com dados internos. Com Big Data podem começar a trabalhar com dados não estruturados que sempre estiveram fora do radar das atividade de BI dos bancos, como conversas efetuadas pelos clientes com o call center, informações geoespaciais obtidas dos smartphones quando operando uma transação com o banco, e atividades dos seus clientes e não clientes nas mídias sociais. Com isso os bancos passam a ter uma visão contextual dos clientes e não apenas o registro das suas operações e transações financeiras. Gerenciamento de riscos é obviamente o primeiro passo, devido à sua importância, bem como pelos regulamentos a que os bancos estão atrelados. Mas com Big Data eles podem fazer análises muito mais refinadas e granuladas, criando vantagens competitivas em relação a outros bancos.

Com Big Data os bancos podem começar sua transição de empresas de serviços financeiros focadas na venda de produtos para uma organização centrada no cliente, oferecendo muito mais serviços personalizados, pelos canais escolhidos pelo próprio cliente. Um exemplo de uso de Big Data em bancos pode ser a utilização de dados obtidos pelas mídias sociais para entender melhor quem é o cliente, o que ele deseja e suas opiniões e sentimentos com relação à marca do banco. Presença nas mídias sociais não é apenas uma fan page no Facebook ou um perfil no Twitter, voltados a divulgar produtos e efetuar campanhas de marketing, mas uma ferramenta de relacionamento e engajamento com seus clientes. Na prática existem milhões de comentários sendo efetuados a cada dia no Facebook e muitos deles podem estar relacionados com experiências, positivas ou negativas de seus clientes com o banco. No Brasil, por exemplo, no final de 2012 um estudo da Socialbakers mostrou que havia quase 65 milhões de usuários ativos mensais, ou seja, um em cada três pessoas no país está no Facebook.

Uma tecnologia que tem muito a colaborar na prevenção de fraudes é a denominada *stream computing, que veremos com maior profundidade na parte 2.* Mas, antecipando a discussão, a

ideia do *stream computing* é fantástica. No modelo de *data mining* tradicional, uma empresa filtra dados dos seus vários sistemas e, após criar um *data warehouse*, dispara "queries". Na prática faz-se garimpagem em cima de dados estáticos, que não refletem o momento, mas sim o contexto de horas, dias ou mesmo semanas atrás. Com *stream computing* esta garimpagem é efetuada em tempo real. Em vez de disparar *queries* em cima de uma base de dados estática, coloca-se uma corrente contínua de dados (streaming data) atravessando um conjunto de queries. Ou seja, com esta tecnologia associada a outras, como ferramentas de análise preditiva e bancos de dados em memória, é possível construirmos sistemas antifraudes bem mais poderosos que os usados atualmente.

Outra área de utilização de Big Data que começa a despontar é a extração do "sentimento" das multidões. É uma área de exploração imensa, pois a interação de muitos indivíduos através das redes sociais conectadas gera enorme e variada quantidade de dados (texto, áudio, imagens estáticas, vídeo etc). Garimpando os dados inseridos nas redes sociais, o "sentimento" das multidões pode ser detectado e torna-se uma das mais ricas fronteiras entre a computação e as ciências sociais. Utilizando técnicas cada vez mais aprimoradas, essas tecnologias tentam "perceber" o que a multidão (milhares de pessoas) está pensando sobre um determinando tema ou fato. Com o propósito de "captar o sentimento" no meio da conversação nas redes, sistemas que reconhecem linguagem e tecnologias com inteligência artificial vasculham semanticamente os posts e tuítes inseridos pelos usuários e analisam seu conteúdo, buscando identificar seu viés com relação ao tema analisado.

Aliás, conhecer a fundo os seus clientes é essencial para as empresas. Uma empresa pode utilizar a analítica comportamental para oferecer aos seus clientes promoções em tempo real e personalizadas por meio da mídia de sua escolha. Estudar e analisar a maneira como os clientes utilizam a Web, o e-mail, o telefone e as redes sociais para pesquisar e comprar permite que a empresa forneça ofertas instantâneas e específicas que gerem resultados positivos. Com Big Data e ferramentas analíticas, as campanhas personalizadas tornaram-se ainda mais precisas e podem ser entregues por meio dos canais mais eficientes para cada cliente. Os responsáveis pelo marketing podem agora ficar mais próximos do que nunca de falar diretamente aos seus clientes sobre os produtos e serviços e de entregar a melhor oferta, automaticamente, com um melhor entendimento de como os clientes interagem e respondem às suas ações. Um exemplo de como explorar o imenso volume de dados de que dispõem é o setor de telecomunicações. Imaginem quanto de informações as operadoras de telefonia têm de seus clientes, que uso fazem dos seus telefones e smartphones, que horários e serviços usam com mais intensidade. Cada cliente das operadoras de telefonia deixa um verdadeiro rastro digital. E coletando e analisando este imenso volume de dados elas podem propor novos serviços de acordo com as demandas específicas de cada cliente. Além disso, passam a ter condições de mapear suas redes, detectando os pontos e horários de maior tráfego, e até de forma preditiva ajustar sua

capacidade antes que surjam problemas de saturação, que afetam a qualidade dos serviços.

Vamos olhar mais de perto este setor. Hoje, em todo o mundo, a cada minuto são enviados 168 milhões de e-mails, seiscentos novos vídeos são levantados para o YouTube, mais de quinhentos mil comentários são registrados no Facebook, quase cem tuítes são enviados e 370 mil telefonemas são feitos pelo Skype. É um imenso volume de dados circulando, mas as operadoras não conseguem usufruir deste potencial todo. É um problema sério. Para dar conta deste crescente volume, as operadoras estão investindo cinco vezes mais em dados que em voz, mas as suas receitas são inversas. Ou seja, faturam cinco vezes mais com voz que com dados. Esta diferença é devido à falta de um modelo de cobrança que esteja alinhado com o perfil de consumo de seus usuários. De maneira geral, os pacotes de dados são vendidos por tamanho e não por tipo de uso. Para cobrar proporcionalmente tanto do usuário comum quanto do intensivo, é necessário integrar dados de sistemas de tarifação com as políticas de controle de uso dos recursos. Com a chegada da quarta geração (4G) a telefonia móvel será totalmente baseada na Internet. O padrão 4G demanda uma convergência muito maior de serviços e será cada vez mais comum o uso de mensagens multímidias (MMS) com vídeo e voz integrados, em vez de simples mensagens de texto. A falta de tecnologia e processos de como tratar dados integrados limita as iniciativas de marketing das empresas de telefonia e com isso elas não conseguem hoje se diferenciar em um mercado cada vez mais competitivo. O uso de Big Data, com capacidade de monitorar e gerenciar transações em tempo real permite personalizar os serviços oferecidos. Por exemplo, imagine monitorar todos os usuários em tempo real. Com esta monitoração a empresa de telefonia pode avisar ao usuário que ele está gastando muita banda e oferecer um pacote de dados extra. Hoje, sem esta monitoração, o usuário não sabe quanto está usando da rede e se ele tem um pacote de vinte megabytes pode ser surpreendido com um aviso de que sua velocidade caiu a 10% do esperado por ter atingido o limite do seu pacote. Ele tem que ligar para o call center e passar por toda a demora que conhecemos para pedir um pacote adicional. Com a monitoração, a operadora avisa a ele antes do problema acontecer.

Educação é um setor que tem muito a ganhar com Big Data. Na verdade, a academia, nos seus centros de pesqusa, já está acostumada a trabalhar com volumes significativos de dados. Basta pensar nos imensos volumes de dados que têm ser analisados em pesquisas que envolvam ramos da ciência como astronomia, química computacional e bioinformática. A computação de alto desempenho e uso de paralelismo massivo é razoavelmente comum no ambiente acadêmico. A crescente digitalização das informações, inclusive dos livros editados em papel desde o início da prensa nos abre espaço para uma nova disciplina, a Culturomics (http://www.culturomics.org/), que é a aplicação de dados em escala massiva para análise e estudo da cultura humana. Um exemplo interessante pode ser visto nesta palestra do TEDx Boston, com o título de “O que aprendemos de cinco milhões de livros”, visto em

http://www.ted.com/talks/what_we_learned_from_5_million_books.html.

Onde Big Data pode ser aplicado na educação? Por exemplo, sugerir, baseado em padrões de milhões de alunos, quais as profissões que melhor se adéquem a cada pessoa. Assim, além dos famosos testes vocacionais, um uso massivo de dados pode ajudar na identificação da melhor combinação personalidade-carreira. O mesmo princípio pode ser adotado para personalizar o estudo, uma vez que nem todos os alunos mantêm o mesmo ritmo de aprendizado e gosto por determinados assuntos. Big Data também pode ser usado em pesquisas que mostrem as variáveis que mais influenciam as taxas de evasão das escolas. As pesquisas podem demonstrar padrões de comportamento que revelem os problemas levando as autoridades responsáveis pela educação a tomar decisões baseadas em fatos e não em subjetividades.

Um uso interessante e ainda pouco comum de Big Data é na própria área de TI. Os setores de TI estão cada vez mais complexos, com profusão de tecnologias de diversas gerações coexistindo, pressões intensas por fazer mais por menores custos e uma crônica escassez de recursos humanos capacitados. A importância da TI nas empresas pode ser medida pelo alto impacto que um sistema fora do ar causa nos negócios. Big Data pode ajudar TI a se autogerenciar. Em muitas empresas, análises estatísticas já são efetuadas em alguns softwares de gerenciamento, como os que fazem a gestão das redes de comunicação. O uso de Big Data amplia este conceito analisando todas as fontes de dados que envolvem TI, inclusive os diversos logs gerados pelas tecnologias. Os softwares de bancos de dados, os sistemas transacionais, as redes, os servidores etc., geram logs que registram suas atividades e uma análise integrada deste imenso volume de dados pode ajudar TI a ser muito mais eficiente, inclusive detectando eventuais problemas antes que eles aconteçam. Imagino que Big Data possa afetar de forma significativa algumas disciplinas de governança de TI propostas pelo modelo ITIL, como gerência de incidentes ou “incident management”, gerência de problemas ou “problem management” e gerência de mudanças ou “change management”.

Outro setor onde Big Data terá papel importantíssimo é na segurança das informações armazenadas digitalmente. Analisando logs, comunicações via e-mails e comentários em mídias sociais é possível antecipar-se a eventuais tentativas de violação de segurança dos sistemas. Aliás, Big Data por si deve gerar preocupações mais amplas quanto a segurança, uma vez que a organização estará acumulando e analisando imensos volumes de dados antes fragmentados e desconexos. A função de CISO (Chief Information Security Officer) pode ser significativamente afetada pelo uso de Big Data nas prevenções e detecções de violações de segurança nos sistemas da corporação. Por exemplo, pode-se detectar correlações nas tentativas de acesso não autorizados, antes não detectáveis, por afetarem sistemas diferentes, como o ERP, o site de comércio eletrônico e mesmo o correio eletrônico. Com cruzamento de informações é possível identificar um determinado padrão e chegar à conclusão de que um eventual infrator está

testando os pontos fracos dos sistemas da empresa. Além disso, com uso de tecnologias de stream computing, as detecções podem disparar ações corretivas e preventivas em tempo real.

O uso de Big Data pelas empresas as ajuda a melhorar diversas ações, antes impensáveis, como:

- **Otimizar o cross-selling ou venda cruzada**. Um bom exemplo é a Amazon, com seus algoritmos de recomendação que, analisando o comportamento e hábitos de leitura dos clientes, recomenda outros livros. Segundo a Amazon, cerca de 30% das suas vendas são provocadas pelas recomendações "you might also want".
- **Location-based marketing ou marketing baseado em localização**. Ao identificar a localização de um cliente, a empresa pode enviar uma mensagem específica a ele com alguma promoção especial. Existe inclusive uma empresa americana chamada PlaceCast que oferece um serviço bem interessante chamado ShopAlerts, que envia mensagens de texto aos clientes de seus clientes (empresas como Starbucks, por exemplo) ao identificar que eles estão em determinada região próxima de uma loja. Segundo a empresa, 79% dos clientes que recebem a mensagem personalizada tornam-se mais propensos a visitar a loja.
- **Análise do comportamento do cliente na loja**. Cerca de 80% do tempo dos clientes dentro de uma loja é gasto circulando na busca por produtos e não efetivamente interagindo ou comprando. Rastrear e analisar o comportamento do cliente na loja permite investir em ações que aproveitem este tempo desperdiçado. Existem start-ups inovadoras que exploram este nicho como a brasileira IDXP e a americana Shopkick.

Big Data abre inúmeras oportunidades para criação de novos negócios. Além das start-ups que citamos, vale a pena estudar o caso da Sense Networks, que é uma empresa com sede em Nova York, fundada em 2003 que se especializou em tratar Big Data e oferecer diversos serviços baseados neste conceito. Uma das suas primeiras aplicações é o CitySense que analisa em tempo real a movimentação em uma cidade e conectando-se ao Yelp e ao Google identifica quais os locais de maior concentração de pessoas. Outra aplicação é CabSense, que, baseado em milhões de informações de GPS identifica qual a esquina mais próxima de você que é adequada naquele momento para pegar um táxi. Outra inovação, que inclusive recebeu o prêmio "Innovation of the Year" da revista americana Popular Science em 2012 é o Google Now. É app para Android, que usa os recursos de Big Data do Google e das tecnologias dos smartphones e tablets para monitorar a rotina de seu usuário, quais sites acessa, que locais frequenta e que tipos de transporte utiliza mais. Com estes dados o app sugere o que o dono do equipamento pode fazer, automaticamente. Por exemplo, ao sair de uma estação de metrô na hora do almoço o usuário pode ser automaticamente avisado dos restaurantes mais próximos e com base nos seus hábitos de frequência de restaurante, os que possam interessar mais.

Capítulo 5

Definindo estratégia de Big Data nas empresas

Estamos apenas no início das descobertas do potencial do Big Data. Nos próximos três a quatro anos veremos as empresas começando a entender e a explorar, embora muitas vezes de forma bastante embrionária os conceitos de Big Data. Sentimos a evolução acontecer de forma rápida. Até fins de 2012 a maioria dos eventos relacionados com o tema tentava explicar "O que é Big Data". Hoje, em 2013, já vemos questionamentos um pouco mais avançados, com empresas buscando respostas para "Como medir o ROI de projetos Big Data" ou "Como fazer Big Data acontecer na minha empresa". É um clara sinalização de que as organizações já começam a olhar Big Data além da simples curiosidade.

A maioria do projetos atuais de Big Data ainda estão na fase exploratória, sendo considerados mais como projetos de prova-de-conceito (Proof-of-concept). Mas é indiscutível que todos os executivos de alto nível, particularmente os CIOs, devem ter uma visão do potencial do Big Data e desenhar uma estratégia adequada para sua adoção. A falta de compreensão do que é Big Data e de seus potenciais e limitações pode gerar riscos para o negócio. Um investimento excessivo em tecnologias sem uma preparação para a empresa explorar seu potencial é jogar dinheiro fora. Se forem extremamente conservadoras e esperarem que o mercado esteja bem maduro antes de iniciar sua jornada de Big Data pode acarretar perda de espaço no mercado. Em resumo, Big Data não pode em nenhuma hipótese ser ignorado.

À medida que Big Data torna-se mais e mais importante para as empresas, seu uso de forma inteligente e inovadora será uma ferramenta de vantagem competitiva inestimável. Como vimos no capítulo anterior, já existem casos de sucesso no uso deste conceito e suas tecnologias. Portanto, adotar Big Data está deixando de ser uma opção, para ser compulsório nas empresas. A questão não é mais se vou ou não adotar Big Data, mas quando e com que estratégia adotarei.

Neste capítulo vamos debater um pouco mais as questões ligadas às estratégias de adoção e uso de Big Data nas empresas.

Antes de mais nada é importante reconhecer que iniciativas de Big Data são diferentes de muitas outras iniciativas de TI. Big Data impacta processos de negócio (pode afetar o processo em tempo real), fontes de dados (começa-se a usar cada vez mais fontes externas à organização), arquitetura de dados e sistemas, infraestrutura e suporte tecnológico (novas tecnologias como bancos de dados NoSQL, por exemplo), estrutura organizacional, e capacitação. Pode afetar de forma drástica a corporação, inclusive mudando o mind set da tomada de decisões baseadas em intuição para fatos. Indiscutivelmente, a complexidade do mundo de negócios atual não permite apenas decisões baseadas em intuição e experiência profissional. Elas continuam valendo, sem dúvida, mas devem ser emparelhadas às análises de fatos, muitas vezes desconhecidos pelos executivos. Além do fato, é claro, que muitas decisões tomadas sob extrema pressão nem sempre são as melhores. Um exemplo: uma grande companhia global identificou gastos excessivos em viagens de seus funcionários. O sistema de aprovação de viagens permitia requisitar a viagem e depois efetuar o reembolso. Com pressão para reduzir custos, ao invés de uma decisão baseada em fatos, como "que funcionários estão gastando em excesso e por quê?" optou-se por criar mais um sistema. Assim, agora é necessária uma prévia autorização, que dependendo do tipo de viagem poderia subir a escalões mais altos da corporação para que fosse autorizada. O resultado foi uma diminuição nos gastos totais de viagem, mas aumentaram muito os custos de cada viagem, pois perderam-se as promoções de vôos comprados com antecedência e indiscutivelmente perderam-se muitas oportunidades de estreitar relacionamentos com clientes. Uma análise apenas de planilhas mostrou os ganhos tangíveis, mas não mostrou as perdas intangíveis.

Big Data tem uma abrangência muito maior que os projetos de BI que as empresas estão acostumadas a desenvolver. BI concentra-se na análise de dados gerados pelos sistemas transacionais enquanto Big Data vai além, explorando fontes de dados externas como comentários e tuítes nas plataformas de mídia social e/ou gerados por sensores e outras fontes geradoras de dados, como RFID acoplados em embalagens e textos gerados a partir das conversas dos clientes com o call center. A diferença é significativa. Os dados transacionais, como os coletados pelos ERP mostram as transações efetuadas, como as vendas de determinados produtos. Armazenando-se o histórico de vendas, podemos fazer análises do comportamento das vendas nos últimos anos e daí tentar extrapolar cenários futuros, mas, sem dúvida, baseados no passado. Com Big Data além das transações efetuadas podemos analisar o que não foi vendido, analisando-se comentários em mídias sociais que abordam por que as pessoas não compraram o produto. Também medindo-se em tempo real os sentimentos do clientes podem-se fazer alterações nas próprias promoções, durante a vigência destas.

No cenário atual, típico do BI, no centro encontram-se os sistemas transacionais e as funcionalidades analíticas giram em torno deles. Em Big Data o contexto é diferente. No centro estão os dados e as capacidades analíticas e em torno destas giram as aplicações. Saímos de um modelo centrado em aplicações para um modelo centrado em analítica.

Com Big Data muda-se o contexto das análises. Em BI geralmente questionam-se coisas como a evolução das vendas durante os últimos anos. Com Big Data entramos em um contexto onde o importante é ter capacidade de gerar novas perguntas. Um exemplo típico seria um questão "como podemos aumentar a fidelização dos nossos clientes em 30%, explorando mais profundamente seus interesses e hábitos de compra, e ao mesmo tempo levando em conta as previsões econômicas e os movimentos da concorrência?". É bem diverso do horizonte limitado do BI atual. Grande parte dos dados são oriundos de fontes externas. Estudos têm mostrado uma crescente utilização de fontes externas para análises mais avançadas. Uma pesquisa com empresas americanas mostrou que em 2011 cerca de 14% das análises usavam bases externas e em 2012 este percentual já havia subido para 31%.

Big Data representa inovação em dois aspectos. Na tecnologia e na forma de processos de tomada de decisões nas empresas. Na tecnologia pois embute bancos de dados NoSQL, processamento massivamente paralelo (comum na academia e centros de pesquisa, mas incomuns no ambiente corporativo) e funcionalidades capazes de coletar, tratar e analisar dados não estruturados como comentários postados no Facebook. Outra inovação é a capacidade do Big Data interferir nos processos da empresa. Para isso acontecer é necessário que os próprios processos sejam revistos para incorporar os resultados das análises nas suas etapas. Um exemplo: ofertas personalizadas para os clientes geralmente são feitas com planejamento e antecedência de vários dias ou semanas. Seleciona-se uma determinada campanha, filtram-se os clientes selecionados e enviam-se para eles e-mails com as ofertas. Com Big Data pode-se identificar uma oportunidade e enviar uma oferta em tempo real.

Tomadas de decisões baseadas em fatos obtidos em tempo real também criam expectativas e receios nos executivos. É importante separar os tipos de decisões tomadas automaticamente. Existem as decisões operacionais que envolvem ações no dia a dia como, por exemplo, um sistema de rastreamento de encomendas que pode identificar, através de uma etiqueta eletrônica (RFID) que o objeto está no rumo errado e tomar decisões de correção imediatas. Está inserido em processos bem-definidos e cercados de precauções com relação aos controles de segurança e acurácia dos dados. Mas existem também decisões automáticas onde eventualmente a situação pode sair do controle, pois envolve uma amplitude maior que um processo fechado. Um sistema de trading, por exemplo, que toma decisões de compra e venda de ações pode, em determinadas situações imprevistas de alta volatilidade do mercado, tomar decisões erradas. Portanto, muitas vezes o elemento humano deve intervir no processo. Não que o humano seja imune a erros, mas,

pode, baseado nos fatos, dar interpretação diferente de um algoritmo, levando em consideração aspectos intangíveis e humanos. Portanto, para aceitarmos tomadas de decisões automáticas temos que ter confiança nos algoritmos e na veracidade dos dados que ele usa para fazer suas análises. Esta confiança é conquistada com o tempo, com o somatório de decisões acertadas.

O primeiro passo para iniciar com Big Data na empresa é identificar os dados que a companhia possui e que pode utilizar. Existem dados internos, que estão nos seus bancos de dados corporativos ou em arquivos de sistemas departamentais, dados que estão em plataformas de mídias sociais e mesmo em bases de dados especializados que podem ser acessados livremente (dados de instituições de governo, públicos) ou adquiridos. Plataformas sociais como Facebook e Twitter são fontes de informação de grande valor, pois alimentam a empresa de opiniões sobre seus produtos e atitudes de forma totalmente livres de censura. É interessante observar que muitas empresas não sabem o valor dos seus dados. Me hospedo em uma rede de hotéis de uma cadeia europeia e, para minha surpresa, ao chegar em um destes hotéis, em uma das capitais brasileiras, o recepcionista me solicitou todos os dados, uma vez que haviam se passado mais de dois meses desde minha última hospedagem. Como ele disse: "apagamos os dados depois de dois meses", fiquei surpreso em ver como a gestão desta rede de hotéis ignora o potencial das informações históricas sobre seus clientes. Existem muitos dados internos que hoje não são analisados. Um exemplo são as conversas gravadas pelas chamadas ao call center que, se armazenadas e processadas como textos, podem ser fontes valiosas de informações sobre os clientes, seus desejos e opiniões sobre a empresa. Muito mais confiáveis que os relatórios escritos pelos atendentes e filtrados pelos seus supervisores. Começar projetos de Big Data usando dados internos tem a vantagem da empresa já dispor destes dados. Faltava tecnologia e a visão de como usá-los.

Aliás, o Gartner está propondo um novo modelo econômico para mensurar o valor das informações, que ele batizou de Infonomics. Infonomics é a disciplina de mensurar e avaliar a significância econômica das informações que uma empresa possui, de modo que estas informações possam ser valorizadas monetária e contabilmente.

É interessante observar que a informação, apesar de todos os discursos sobre seu valor competitivo, não é valorizada monetariamente. Por exemplo, se um data center pegar fogo, as seguradoras cobrem o prejuízo sofrido pelas instalações e pelo maquinário, de geradores a servidores. Mas não cobre o conteúdo dos dados que foram perdidos. De maneira geral, uma empresa com boa governança de TI mantém uma política de backup eficiente e consegue recuperar todas ou quase todas as informações. Mas caso não consiga, ela não obterá da seguradora a reparação pelo valor dos dados perdidos, pois estes não são valorizados monetária e contabilmente.

Vivemos hoje na sociedade da informação e informação é um produto por si mesmo, além de ser o combustível que impulsiona os negócios da maioria das empresas. A consequência deste fato é o surgimento de tecnologias de bancos de dados, data warehouse e mais recentemente o próprio conceito de Big Data.

Se analisarmos a informação vemos que ela se encaixa perfeitamente nas características de um bem econômico intangível, que são:

a. **Custo relativamente alto para sua criação**. A produção da informação custa muito mais que as cópias geradas, que têm custo marginal.
b. **Escalabilidade**. Custos marginais para produzir duas ou duas centenas de cópias. Atualmente com armazenamento e cópias inteiramente digitais elimina-se também o custo de produção das cópias impressas.
c. **Economias de escala em termos de produção**. No caso da informação impressa, como em livros, quanto maior a edição, menores os custos individuais devido à economia de escala. Nos meios digitais, como e-books, tais custos inexistem.
d. **Pode ser usada por mais de uma pessoa a cada momento**. Diferente de um bem tangível como um carro, que se eu estiver dirigindo ninguém mais poderá dirigi-lo ao mesmo tempo.
e. **Substituição imperfeita**. Uma cópia reduzida em conteúdo ou fragmentada não pode substituir a informação completa, original.
f. **Efeito de rede**, cujo valor cresce à medida que mais pessoas a utilizam.

O modelo de Infonomics propõe valorizar a informação. Isto significa quantificar de forma tangível a informação, de modo que possamos dizer que esta determinada informação vale 350 mil reais e esta outra quinhentos mil reais. Isto significa que ela poderá ser incluída nas análises contábeis e fazer parte do valor de uma empresa. Uma empresa que usar mais inteligentemente suas informações que outras será mais bem avaliada em termos de valor de mercado. Numa comparação simples com empresas da sociedade industrial, como uma companhia de petróleo, vemos que o valor desta é estimado pelos repositórios de petróleo de que ela dispõe (suas reservas) bem como pela sua capacidade de extrair e refinar este petróleo. Levando para o conceito do Infonomics, uma empresa será valorizada pelo valor das informações que contém e pela sua capacidade de explorá-las adequadamente. Este é um ponto interessante. Informação, mesmo que não seja usada, tem seu valor. Assim como uma mercadoria em centro de distribuição tem seu valor (valor do estoque) antes mesmo de ser vendida, a informação tem valor, mesmo antes de ser tratada por tecnologias de análise de dados. Podemos começar a medir seu valor potencial.

Um benefício desta valorização é que torna mais simples a proposição de projetos que

envolvam os conceitos de manuseio de informação, como Big Data. Será possível, com Infonomics conseguir mostrar que determinado projeto aumentará em 100% o valor de determinada informação, facilitando gerar as estimativas de ROI (Retorno sobre investimento) destes projetos.

Um detalhamento mais aprofundado do assunto Infonomics e links para outras fontes pode ser acessado em http://en.wikipedia.org/wiki/Infonomics.

O segundo passo na estratégia de Big Data é identificar as oportunidades de explorar as informações disponíveis, levantadas na etapa anterior. Uma sugestão é começar usando o potencial de Big Data por setores de alto potencial, como marketing digital e a partir destas primeiras experiências, replicar o feito por toda a organização. É um processo evolutivo, que depende do grau de maturidade de gestão da empresa. Não acontece de um dia para outro, ou seja, não existe Big Bang para Big Data.

O terceiro passo é criar uma infraestrutura organizacional e de processos para aproveitar de forma adequada os insights gerados pela exploração do Big Data. Além de profissionais capacitados para fazerem análises e mesmo perguntas aos sistemas, é importante que as informações geradas provoquem reações na empresa. Para isso os processos devem considerar os resultados das análises. É importante disseminar a cultura de Big Data e da importância de análises nos gestores da organização. Na verdade propomos criar uma mind set "data-driven" ou uma cultura orientada a dados e informações embutidas no próprio DNA da organização. Um exemplo interessante de influência do DNA de Big Data na organização é a lenda urbana que diz que Jeff Bezos, CEO da Amazon demitiu toda uma equipe de webdesigners que alteraram o web site da Amazon.com porque não fizeram análises exaustivas de reações e comportamentos dos clientes com relação às mudanças propostas. Nos casos onde o sistema de Big Data atua em tempo real, com o uso de tecnologias como stream computing, os processos devem ser ajustados para permitirem que o próprio sistema faça os ajustes automaticamente.

O quarto passo é desenhar uma estratégia de tecnologia para inserir Big Data na organização. Isto implica em ter uma visão holística e integrada dos modelos de dados que permeiem toda a organização, bem como de aquisição de tecnologias adequadas para as iniciativas de Big Data. Como o fenômeno Big Data é bem recente, muitas empresas estão embarcando em iniciativas sem uma estratégia bem-definida. Big Data não pode ser visto como ação isolada, um projeto isolado em alguma área da empresa. Uma questão importante e que afeta os gestores de TI é que, na maioria dos casos atuais de projetos de Big Data, as iniciativas começam pelas linhas de negócio e não por TI. Entretanto, devido à necessidade de integração de tecnologias e acessos a inúmeras bases de dados corporativos, e das questões de segurança e privacidade, TI deve atuar de forma proativa. Caso contrário, terá uma bomba-relógio em suas mãos! O risco

da falta de planejamento é investir muito sem obter resultados ou investir pouco e perder as oportunidades para os concorrentes.

A estratégia de tecnologia começa com uma análise do portfólio tecnológico de que a empresa dispõe para fazer uso do conceito de Big Data. Muitas das tecnologias tradicionais em uso hoje não são de todo adequadas para iniciativas que envolvam imensos volumes de dados, ampla variedade e que respondam na velocidade adequada. Imaginem uma situação. Você tira uma foto de uma avenida bem movimentada há cinco minutos atrás. Agora, você a atravessaria baseado exclusivamente nas informações desta foto? Claro que não, pois o momento presente terá um outro fluxo de veículos e o caminhão que o atropelará, ao ser tirada a foto, estava a quilômetros de distância. A maioria das tecnologias de que a empresa dispõe são baseadas neste cenário, de armazenar informações passadas e de lá periodicamente extrair relatórios e gráficos e de posse deles fazer estimativas futuras. As estruturas tradicionais de data warehouse, sejam processos e tecnologias, não se adaptam a respostas em tempo real.

O modelo de evolução de Big Data na empresa começa com um primeiro nível, no qual muitas empresas já estão, que é extrair relatórios de análises descritivas (Businesss Intelligence tradicional) em cima do data warehouse. A pergunta que se faz é “O que aconteceu?”, uma vez que são análises em cima de dados históricos. O segundo nível de maturidade é quando a empresa começa a trabalhar os dados em tempo real e consegue responder a questões do tipo “o que está acontecendo agora?”. O próximo nível, onde existem apenas alguns poucos casos de sucesso ainda, é quando a empresa entra na fase de análises preditivas e responde a perguntas do tipo “o que acontecerá?”. Finalmente, o último estágio de evolução do Big Data é quando conseguimos chegar à análise prescritiva e respondermos a “como podemos fazer isso acontecer?”.

Um aspecto importante é criar um nível de capacitação adequada. Os skills necessários são para executar tarefas de integrar e preparar dados, modelar e analisar estes dados e suportar as tecnologias envolvidas. Atualmente há escassez de pessoal qualificado para atuar com Big Data. Indo além, provavelmente será necessário repensar a organização de TI. À medida que os usuários reconhecem valor em explorar Big Data, mais e mais pressões para incrementar seu uso cairão na área de TI. De maneira geral, a maioria das áreas de TI estão dando, hoje, mais atenção aos projetos considerados “core” que a projetos especulativos e experimentais de Big Data. Isto pode provocar ações dos usuários buscando por outros meios, usando inclusive as disponibilidades e facilidades de nuvens públicas, soluções para conseguir seus projetos, passando por cima da área de TI. Para evitar problemas futuros, principalmente relacionados com aspectos de segurança e privacidade, é importante que a área de TI não fique alheia a estes movimentos, mas que aja de forma proativa, buscando apoiar estes projetos. O modelo ideal é TI desenhar a arquitetura de Big Data, deixando os usuários criarem seus próprios modelos de

busca e análises por processos self-service. TI fica responsável pela infraestrutura, em casa ou em nuvem pública, integração e validação dos dados e pela governança, de maneira geral.

Um outro ponto importante é que à medida que a empresa amadurece no uso do Big Data, cria uma nova posição, a de CDO (Chief Data Officer) que com sua equipe de cientistas de dados não se posiciona de forma subordinada ao CIO, mas sim ligado diretamente ao CEO ou ao CMO (Chief Marketing Officer). Portanto, tanto o CDO como o CIO deverão criar forte e estreito relacionamento colaborativo pois o trabalho de ambos é inter-relacionado. Desta forma, o CDO mantém a equipe de cientistas de dados e pessoal capacitado a fazer as análises avançadas e o CIO se encarrega de manter a equipe técnica que suporta o ambiente tecnológico de Big Data.

Finalmente, é importante definir uma política de uso das informações, considerando os aspectos de privacidade e segurança. Privacidade implica em aderências às legislações e aspectos regulatórios que estejam em vigor. Também é importante manter um acompanhamento constante pelo fato de Big Data ainda ser uma novidade e muita legislação nova deve ser gerada nos próximos anos. Muitas discussões sobre privacidade sairão na mídia e através de ONGs preocupadas com estas questões. A segurança também deve ser vista com atenção, pois se a empresa consegue acumular muitas informações sobre seus clientes, como comportamento, hábitos etc., deve tomar muito cuidado para que tais informações não sejam acessadas e divulgadas indevidamente. Mesmo quando informações individualmente não contenham explicitamente informação sobre determinado indivíduo, quando triangulada e integrada com diversas outras pode expor segredos industriais ou identificação pessoal íntima. Esta situação pode gerar não apenas processos judiciais, mas arranhar significativamente a imagem da corporação. Portanto, é essencial criar uma política de governança para dados e informações no mundo do Big Data.

PARTE II

A tecnologia por trás do Big Data

Capítulo 6

A infraestrutura de tecnologia

Na parte 1 analisamos Big Data sob a ótica das mudanças e impactos que ocasionará nos negócios e na sociedade. Nesta segunda parte vamos abordar o tema pela ótica da tecnologia.

O termo Big Data está cada vez mais popular, embora ainda esteja mal-compreendido. Observo que ainda não existe consenso quanto ao que realmente é Big Data e quais as tecnologias fundamentais que o sustentam. E mais ainda, existem muitas dúvidas de como tangibilizar o conceito, ou seja, como sair do conceitual e criar soluções de negócio que agreguem valor para as companhias.

Eliminar estas dúvidas é essencial e o primeiro passo para as empresas se aventurarem em projetos Big Data.

Para colocarmos o termo em contexto, Big Data vem chamando atenção pela acelerada escala em que volumes cada vez maiores de dados são criados pela sociedade. Já falamos comumente em petabytes de dados gerados cada dia e zetabytes começa a ser uma escala real e não mais imaginária e futurista. O que era futuro há uma década, terabytes, hoje nós já temos nas nossas próprias casas.

As tecnologias que sustentam Big Data podem ser analisadas sob duas óticas: as envolvidas com analytics, tendo Hadoop e MapReduce como nomes principais e as tecnologias de infraestrutura, que armazenam e processam os petabytes de dados. Neste aspecto, destacam-se os bancos de dados NoSQL (No, significa not only SQL). Por que estas tecnologias? Por que Big Data é a simples constatação prática de que o imenso volume de dados gerados a cada dia excede a capacidade das tecnologias atuais de os tratarem adequadamente.

Vamos relembrar o que é Big Data através da fórmula que mostramos no capítulo anterior: Big Data = volume + variedade + velocidade + veracidade gerando + valor.

Volume está claro. Geramos petabytes de dados a cada dia. E estima-se que este volume dobre a cada dezoito meses. Variedade também, pois estes dados vêm de sistemas estruturados (hoje minoria) e não estruturados (a imensa maioria), gerados por e-mails, mídias sociais (Facebook, Twitter, YouTube e outros), documentos eletrônicos, apresentações estilo Powerpoint, mensagens instantâneas, sensores, etiquetas RFID, câmeras de vídeo etc. Velocidade porque muitas vezes precisamos agir praticamente em tempo real sobre este imenso volume de dados, como em um controle automático de tráfego nas ruas. Velocidade é um critério que vai se tornar cada vez mais importante, devido à crescente rapidez com que as empresas precisam reagir às mudanças no cenário de negócios, bem como é necessária para tratar os dados em tempo real, interferindo na execução do próprio processo de negócios. Veracidade porque precisamos ter certeza de que os dados fazem sentido e são autênticos. E valor porque é absolutamente necessário que a organização que implementa projetos de Big Data obtenha retorno destes investimentos. Um exemplo poderia ser a área de seguros, onde a análise de fraudes poderia ser imensamente melhorada, minimizando-se os riscos, utilizando-se, por exemplo, de análise de dados que estão fora das bases estruturadas das seguradoras, como os dados que estão circulando diariamente nas mídias sociais.

Entretanto, observamos que as tecnologias atuais de tratamento de dados não são mais adequadas para atender a esta nova demanda provocada pelo conceito de Big Data. Por quê? Vejamos o modelo relacional, proposto pelo pesquisador da IBM, Edgar F. Codd, em 1969. Quando foi proposto, a demanda era acessar dados estruturados, gerados pelos sistemas internos das corporações. Não foi desenhado para dados não estruturados (futurologia na época) e nem para volumes na casa dos petabytes de dados (inimaginável na época). Precisava-se, sim, de um modelo que categorizasse e normalizasse dados com facilidade. E o modelo relacional foi muito bem-sucedido nisso, tanto que é o modelo de dados mais usado atualmente.

Claro que existem ainda grandes desafios pela frente. Um deles é a evolução da tecnologia para manusear rapidamente este imenso volume de dados. Existem algumas tecnologias orientadas a tratar volumes muito grandes como Hadoop e sistemas de bancos de dados específicos como o Cassandra, sistema Open Source (http://cassandra.apache.org/) utilizado hoje pelo Facebook, Twitter e Reddit que precisam tratar com muita velocidade imensos volumes de dados de forma distribuída.

Para tratar dados na escala de volume, variedade e velocidade do Big Data precisamos de outros modelos. Surgem os softwares de banco de dados NoSQL, desenhados para tratar imensos volumes de dados estruturados e não estruturados. Existem diversos modelos de

sistemas colunares como o Big Table, usado internamente pelo Google (é a base de dados sob o Google App Engine) e que pode ser visto em http://static.googleusercontent.com/external_content/untrusted_dlcp/research.google.com/pt-BR//archive/bigtable-osdi06.pdf,o modelo Key/value como DynamoDB da Amazon (http://aws.amazon.com/pt/dynamodb/), o modelo "document database" baseado no conceito proposto pelo Lotus Notes da IBM e aplicado em softwares como MongoDB, e o modelo baseado em grafos como o Neo4j (http://neo4j.org/). Em resumo, não faltam opções. Interessante lembrar que antes do modelo relacional já existia um software de banco dados que lidava com grandes volumes que é o IMS da IBM, modelo hierárquico, criado para suportar o projeto Apollo de conquista da Lua e que ainda hoje é base da maioria das transações financeiras que circulam pelo mundo. Mais detalhes do IMS em http://www-01.ibm.com/software/data/ims/.

Por outro lado, esta diversidade de alternativas demanda que os líderes dos projetos de Big Data escolham a mais adequada ou mesmo usem mais de uma opção, de acordo com as necessidades específicas.

Depois da infraestrutura é necessário atenção aos componentes de analytics, pois estes é que transformam os dados em algo de valor para o negócio. Big Data Analytics não significa eliminar os tradicionais sistemas de BI que existem hoje, mas, pelo contrário, devem coexistir. Recomendo enfaticamente a leitura do livro "Competing on Analytics: the new science of winning", de Thomas H. Davenport, publicado pela Harvard Business Schoool Press. Um bom exemplo de uso de Hadoop para analytics é o BigInsights da IBM, em http://www-01.ibm.com/software/data/infosphere/biginsights/.

Quanto ao aspecto velocidade, o conceito de stream processing permite tratamento em tempo real de dados. Concretamente, o InfoSphere Streams da IBM é um exemplo muito interessante. Vamos analisá-lo no capítulo seguinte.

Além das tecnologias de tratamento analítico de dados, são necessárias evoluções significativas na maneira de se visualizarem os dados. É um campo que tem demandado muita pesquisa. Existem resultados interessantes como "tag cloud" (http://en.wikipedia.org/wiki/Tag_cloud), clustergramas (http://www.schonlau.net/publication/02stata_clustergram.pdf), History Flow (http://www.research.ibm.com/visual/projects/history_flow/) e Spatial Information Flow, onde um belo exemplo pode ser encontrado em http://senseable.mit.edu/nyte/.

Adicionalmente, podemos pensar que a computação em nuvem é também um impulsionador para Big Data, pois podem-se usar nuvens públicas para suportar imensos volumes de dados e as características de elasticidade das nuvens permitem que acionemos servidores virtuais sob

demanda, apenas no momento de tratar estes dados. Na prática, o modelo de cloud computing faz muito sentido para as atividades de Big Data, dada a imprevisibilidade do montante de dados que poderá ser tratado. O ajuste da capacidade de armazenamento e processamento varia imensamente e portanto torna-se muito difícil planejar-se, *a priori*, uma configuração exata. Usar o conceito tradicional de configurar sistemas pelo máximo torna-se inviável devido a esta imprevisibilidade. O fator elasticidade é um parâmetro importante no projeto da infraestrutura de Big Data. Além disso, o custo tende a ser bem menor que montar uma infraestrutura tradicional, com servidores, armazenamento para grandes volumes de dados e softwares. A Amazon, por exemplo, lançou um serviço chamado Redshift que, na prática, é um data-warehouse-as-a-service cobrando, quando da escrita deste livro, mil dólares por petabyte por ano. Com estes valores, empresas de pequeno porte podem usufruir do potencial do Big Data. Provavelmente veremos em breve muitas ofertas de BigData-as-a-Service, ofertadas por provedores de nuvem.

Um caso interessante de uso de Big Data em nuvem é o da Etsy, site de e-commerce especializado em produtos de artesanato e artigos de época, que já tem mais de onze milhões de usuários, resultando em 25 milhões de visitantes únicos e 1,1 bilhões de page views por mês. Hoje Etsy captura mais de 5GB de dados por dia. Este imenso volume de dados é analisado em uma nuvem pública para entender melhor o comportamento dos seus clientes e realizar análises preditivas. Assim, o Etsy consegue determinar quais os produtos que melhor se adequam ao gosto de um determinado cliente. E não precisa instalar grandes servidores para fazer esta análise. A Etsy utiliza a nuvem e paga apenas pelo tempo necessário para realizar a tarefa.

Uma outra variável que se torna cada vez mais importante é a resiliência. Quando começamos a tratar dados em tempo real, ou seja, a tempo de influir na execução de um processo, a infraestrutura de Big Data tem que ser desenhada para alta resiliência e disponibilidade. Um exemplo simples mostra a importância deste fator: suponhamos que estejamos analisando dados em tempo real das condições de trânsito, de modo a impedir congestionamentos. Se a infraestrutura não estiver disponível, o impacto da falta de informações será refletido em engarrafamentos que afetarão em muito a qualidade de vida dos cidadãos.

Big Data vai causar significativos impactos no setor de TI. Não se trata apenas de volumes massivos de dados a serem tratados, mas todas as outras variáveis como velocidade e variedade devem ser compreendidas e previstas por TI. TI deve planejar sua arquitetura tecnológica para suportar esta demanda, seja em equipamentos próprios ou, cada vez mais provável, em ambiente de computação em nuvem. Por exemplo, para obter velocidade será necessário o uso de novas tecnologias como stream computing e bancos de dados residentes em memória. Também precisa-se dispor de tecnologias que coletem dados em diversas fontes,

sejam mídias sociais, câmeras de vídeo ou sensores. Para garantir que os dados sejam utilizáveis é importante dispor de tecnologia que consiga eliminar ruídos e sujeiras dos imensos volumes de dados coletados. Isto tudo implica em diversos formatos e padrões de dados e desenhos de arquitetura de dados que contemplem toda esta complexidade. Acabaram-se os tempos onde o modelo relacional era praticamente o padrão de dados da empresa. Os arquitetos e os sistemas devem estar preparados para trabalhar com diversos formatos de dados, de relacional a NoSQL, bem como integrar tecnologias legadas equipadas com sistemas que usam bancos de dados relacionais a novos sistemas que usarão tecnologias ainda estranhas à TI, como Hadoop.

Claro que nem todas as aplicações serão afetadas por Big Data. Mas a maioria será afetada, de alguma forma. Portanto, é importante que o conceito seja considerado na arquitetura de dados e sistemas das empresas. Na prática, com a proliferação de dispositivos móveis, as plataformas de mídia social e a Internet das Coisas, o volume de dados gerados que pode ser de interesse para as empresas crescerem assustadoramente. Muitas informações vitais podem estar circulando a cada minuto e eventualmente poderão interferir nos processos de negócio, em tempo real.

Entretanto, a imensa maioria dos sistemas atuais das empresas não foi desenhada para este modelo. Concentram-se em modelos de dados estruturados e não utilizam tecnologias como processamento massivamente paralelo ou stream computing.

Mas, os novos sistemas devem considerar o conceito de Big Data e isto implica em repensar a arquitetura de sistemas e dados da corporação. Modelos de dados não estruturados (NoSQL) e relacionais devem fazer parte da arquitetura. O uso de computação paralela e grandes volumes de dados também devem fazer parte da arquitetura. Por exemplo, quando usando stream computing é importante definir que dados serão processados localmente nos sensores e dispositivos remotos e quais serão enviados para processamento nos algoritmos que rodam nos clusters. Além disso, alguns dados coletados têm sentido se aplicados apenas em tempo real, perdendo sua importância com o tempo. Neste caso, é provável que apenas um resumo seja armazenado para eventuais análises futuras.

Capítulo 7

Hadoop

Uma tecnologia que se destaca no cenário de Big Data é o Hadoop. E, portanto, vamos dedicar a ela um pouco mais de atenção. O Hadoop é um projeto da comunidade Apache (//hadoop.apache.org), foi criado pelo Yahoo em 2005, inspirado no trabalho do Google em seu GFS (Google File System) e no paradigma de programação MapReduce, que basicamente divide o trabalho em tarefas como um mapeador (mapper) e um resumidor (reducer) que manipulam dados distribuídos em um cluster de servidores usados de forma massivamente paralela.

Hoje vem, de forma crescente, sendo adotado por empresas que precisam tratar volumes massivos de dados não estruturados. Já existe inclusive um ecossistema ao seu redor, mas ainda vai demandar algum tempo para se disseminar de forma mais ampla pelo mercado. Neste capítulo vamos debater um pouco mais o que é e o que não é o Hadoop, seu mercado e tentar visualizar algumas tendências. Quem sabe acertamos algumas?

Mas, o que é o Hadoop? É, na prática, uma combinação de dois projetos separados, que são o Hadoop MapReduce (HMR), que é um framework para processamento paralelo e o Hadoop Distributed File System (HDFS). O HMR é um spinoff do MapReduce, software que o Google usa para acelerar as pesquisas endereçadas ao seu buscador. O HDFS é um sistema de arquivos distribuídos otimizados para atuar em dados não estruturados e é também baseado na tecnologia do Google, neste caso o Google File System. Existe também o Hadoop Common, conjunto de bibliotecas e utilitários que suportam os projetos Hadoop. Na prática, para que o HMR processe os dados, eles devem estar armazenados no HDFS.

O HDFS (Hadoop Distributed File System) é o sistema que armazena os dados para o Hadoop. Os dados no HDFS são divididos em pequenos pedaços, chamados de blocos e distribuídos por diversos servidores. Desta forma, o processamento subsequente é muito

acelerado pois, em vez de uma pesquisa sequencial, os dados são pesquisados de forma simultânea, em paralelo. A proposta do Haddop é usar servidores e seus discos locais, e como concepção de projeto, considera o uso de servidores e discos baratos. Isto implica que o projeto considera que o MTBF (Mean Time Between Failures ou tempo médio entre falhas) dos equipamentos do cluster não é necessariamente muito alto e portanto o próprio Hadoop deve investir em tecnologias que garantam a disponibilidade do acesso. Portanto, cada bloco é copiado em dois outros lugares diferentes, de modo que se um ou dois servidores falharem, um terceiro supre as necessidades da pesquisa. Os blocos são tipicamente de 64 MB, o que é muito maior que os blocos típicos dos bancos de dados relacionais, que geralmente são de 4 Kb a 32 KB. Lembramos que o Hadoop foi projetado para manusear imensos volumes de dados e faz sentido ter blocos grandes, pois assim cada servidor no processamento paralelo pode trabalhar com um número bastante razoável de dados. A tarefa de coordenar clusters de servidores provoca um overhead significante e o uso de grandes blocos de dados diminui sensivelmente a demanda de informações que precisam ser trocadas entre os servidores.

A lógica de mapeamento dos blocos de dados fica a cargo de um servidor especial chamado NameNode. Para aumentar o desempenho, toda a informação de mapeamento do NameNode é mantida em memória. Mas, chamamos a atenção para um detalhe: só existe um único NameNode, o que cria uma situação que denominamos de Single Point of Failure (SPOF ou um único ponto de falha), que torna vulnerável todo o sistema. Portanto, os projetos de Hadoop enfatizam que o servidor do NameNode seja um servidor de alta disponibilidade e que seja definido um processo de backup/recovery altamente eficiente. A perda do NameNode implica na perda de acesso a todos os blocos, ou seja, todos os dados do Hadoop! A partir da versão 0.21 foi implementada uma nova funcionalidade, BackupNode que opera como servidor standby do NameNode entrando em ação se houver falha no servidor do NameNode.

O outro componente do Hadoop é o MapReduce. É o coração do Hadoop. É o paradigma de programação que possibilita escalabilidade massivamente paralela em centenas ou milhares de servidores. O próprio termo MapReduce representa duas tarefas distintas que os programas Hadoop executam. A primeira tarefa é mapear os dados, ou seja, acessar um conjunto de dados e convertê-los em outro conjunto onde os elementos individuais são quebrados em tuplas (pares chave/valor). A tarefa Reduce pega o resultado do mapeamento e combina estas tuplas em um conjunto menor de tuplas, obtendo o resultado.

O Hadoop é um projeto Open Source, com licenciamento Apache e portanto permite a criação de um ecossistema de negócios baseados em distribuições específicas. E o surgimento de serviços em nuvem, como o Amazon Elastic MapReduce, permite às empresas tratarem dados massivos sem demandar aquisição de servidores físicos. Neste modelo, o usuário escreve a aplicação Hadoop e a roda em cima da nuvem da Amazon.

A base das distribuições Hadoop é a comunidade Apache. Diversas empresas vêm contribuindo com código para seu desenvolvimento como Yahoo, Facebook, Cloudera, IBM e outras. Em torno do código base, surgem diversas distribuições, como Cloudera (www.cloudera.com) e DataStax (http://www.datastax.com/brisk), que agregam valor com utilitários e serviços de suporte e educação, no mesmo modelo das distribuições Linux. A distribuição da DataStax, chamada de Brisk, substituiu o HDFS por um sistema de arquivos distribuídos baseados no software NoSQL Cassandra, chamado agora de CassandraFS. Algumas outras distribuições são a Amazon (Amazon Elastic MapReduce), IBM (InfoSphere BigInsights), EMC (Greenplum HD Community Edition) e Pentaho.

A proliferação do uso do Hadoop abre oportunidades para diversas empresas se posicionarem como distribuidoras. Entretanto, apesar do código base ser o mesmo, cada distribuição tem características diferenciadas e é importante avaliá-las sob critérios de análise rigorosos. Estes critérios podem, por exemplo, ser estruturados em três grupos, considerando sempre o atendimento ao mercado brasileiro:

a. **Oferta atual**. Como a distribuição opera hoje, a arquitetura e funcionalidades Hadoop embutidas, o nível de integração dos subprojetos, funcionalidades de modelagem, suporte a baixa latência e alta performance (lembrando que o Hadoop é voltado a processos batch), cluster management, conectores com aplicativos e softwares de mercado. Validar também as condições de suporte, educação e consultoria disponíveis. Verificar a quantidade e expertise do seu pessoal técnico.
b. **Estratégia**. Qual o direcionamento estratégico da distribuição. Validar também a importância do Hadoop no portfólio da empresa. Algumas distribuições têm como único negócio o Hadoop e outras o têm como mais um produto em seu portfólio. Neste último caso, validar a importância que a distribuição Hadoop tem na estratégia e no volume de receitas da empresa.
c. **Presença no mercado**. Validar o market-share e o ecossistema criado em torno da distribuição. Analisar volume de receita e validar se há condições da distribuidora se manter operando. Como sugestão recomendo solicitar pelo menos dois casos reais de uso do Hadoop da distribuição.

Em torno do Hadoop (http://hadoop.apache.org/), a comunidade Apache mantém diversos projetos relacionados, como o Hbase, que é um banco de dados NoSQL que trabalha em cima do HDFS. Este banco de dados é usado pelo Facebook para suportar seu sistema de mensagens e os seus serviços de informações analíticas em tempo real. Existe também o Hive, criado pelo Facebook, que é uma camada de data warehouse que roda em cima do Hadoop. Utiliza uma linguagem chamada Hive SQL, similar à SQL, o que facilita sua utilização, pois desenvolvedores acostumados com SQL não encontram maiores dificuldades em trabalhar com

o Hive SQL.

Um outro e também muito interessante projeto é o Pig, criado pelo Yahoo. É uma plataforma que permite análises de arquivos muito grandes usando uma linguagem de alto nível chamada de Pig Latin. Olhando-se o stack de softwares do Hadoop, o Pig se situa entre o Hive e o HMR e é uma tentativa de fornecer uma linguagem de alto nível para se trabalhar com o Hadoop. Outros projetos menos conhecidos são o Avro (sistema de serialização de dados), o Chukwa (monitoramento de sistemas distribuídos) e o Hama (para computações científicas massivas).

O quadro a seguir mostra as camadas funcionais do Hadoop e os seus principais subprojetos:

Camada funcional do Hadoop	Subprojetos
Modelagem e desenvolvimento	MapReduce, Pig, Mahout
Armazenamento e gestão de dados	HDFS, Hbase, Cassandra
Data Warehousing e queries	Hive, Sqoop
Coleta, agregação e análise de dados	Chukwa, Flume
Metadados, tabela e esquemas	HCatalog
Cluster management, job scheduling e workflow	Zookeeper, Oozie, Ambari
Serialização de dados	Avro

A IBM, por exemplo, usa intensamente o Hadoop em diversos projetos, o integrando com outros de seus softwares como o Cognos, criando soluções para tratamento analítico de dados massivos e não estruturados, como o InfoSphere BigInsights, que agrega um conjunto de tecnologias open source como o próprio Hadoop, Nutch e Pig, com as tecnologias próprias da

IBM como InfoSphere e ManyEyes. Vejam em (http://www-01.ibm.com/software/data/bigdata/). A IBM também desenvolveu uma variante do HDFS chamado de IBM General Parallel File System (GPFS), que pode ser visto em http://www-03.ibm.com/systems/software/gpfs/.

Como vimos, Hadoop e seu ecossistema de softwares é Open Source. O modelo Open Source já não é mais desconhecido e está embutido hoje na maioria das empresas. Um exemplo é o Linux, que é lugar comum nos servidores e é base da maior parte dos smartphones via Android. A imensa maioria dos serviços Web é motorizada por Open Source, como o Google, a Amazon, o Facebook (http://www.facebook.com/Engineering) e o Instagram. Para este último, como exemplo, vale a pena dar uma olhada no seu stack de software, no artigo "What Powers Instagram" em http://instagram-engineering.tumblr.com/post/13649370142/what-powers-instagram-hundreds-of-instances-dozens-of. Mas, não são apenas empresas ícones do mundo Web que usam Open Source. A IBM, símbolo do mundo corporativo suporta intensamente Open Source e vale a pena dar uma olhada em http://www-03.ibm.com/linux/ossstds/oss/ossindex.html.

Desde que comecei a me interessar mais por Open Source, e isto vai lá pelos meados da década de 90, acompanhei a sua evolução, principalmente no tocante ao amadurecimento do mercado. Escrevi aqui pela Brasport, em 2004 o livro Software Livre: Potencialidades e Modelos de Negócio e o meu foco nestes anos todos foi principalmente avaliar como Open Source impactou a indústria de software, tanto fabricantes como usuários. Um livro que me chamou muito a atenção e que é, no meu entender, a base conceitual do Open Source é "The Cathedral and the Bazaar", de Eric Raymond (http://en.wikipedia.org/wiki/The_Cathedral_and_the_Bazaar). Eric mostrou que o Open Source é um novo método de desenvolvimento de software, colaborativo, que denominou de bazar, pois é constituído de uma comunidade de desenvolvedores voluntários, que atuam no projeto sob suas próprias agendas e vontades. Um bazar, que ele usou como referência é algo assim, meio zoneado. O seu contraponto é o modelo tradicional, a catedral, onde o software é desenvolvido por uma equipe fechada e gerenciada de forma hierárquica, com cronogramas rígidos e com os desenvolvedores atuando nas partes dos códigos que lhe foram designadas e sem escolher o que e quando desenvolver.

O modelo Open Source se mostrou plenamente viável e o Linux e centenas de outros projetos estão aí para confirmar isso, apesar de muitas críticas e oposições iniciais. Portanto, de meados dos anos 90 para cá evoluiu-se muito. O mercado amadureceu, os temores quanto à viabilidade do Open Source desapareceram ou foram muito minimizados e a indústria de software continuou forte. Saímos também de uma discussão ideológica do bem contra o mal e chegamos ao consenso de que radicalismos não nos levam a lugar nenhum. Open Source e softwares

comerciais estão aí, nas empresas, nas casas e agora nos smartphones e tablets.

OK, e quem usa Hadoop? Existem os casos emblemáticos como Facebook, Yahoo, Twitter e Netflix (na nuvem da Amazon), mas também já começamos a ver seu uso em ambientes corporativos brick-and-mortar. Recentemente uma pesquisa mostrou que pelo menos umas vinte empresas da lista da Fortune 1000 assumiram publicamente que usam Hadoop de alguma forma. A adoção do Hadoop em aplicações analíticas corporativas como as ofertadas pela IBM vão ajudar na sua disseminação. Só para lembrar, quando a IBM anunciou seu apoio ao Linux, em 2001, o Linux passou a ser visto sob outra ótica pelo ambiente corporativo.

Nem todo usuário de Hadoop demanda uma escala massiva de dados ao nível do Facebook ou Yahoo. Mas, empresas com razoável volume de informações não estruturadas, como bancos, varejo, empresas aéreas e outras vão encontrar no Hadoop uma boa alternativa para o tratamento analítico dos seus dados. Vejam alguns exemplos em ftp://public.dhe.ibm.com/common/ssi/ecm/en/imb14103usen/IMB14103USEN.PDF.

O Hadoop ainda está nos primeiros passos de sua evolução e disseminação pelo mercado. E como tal, apresenta os problemas da imaturidade. Por exemplo, até o surgimento da versão 1.0.0 em dezembro de 2011, não havia um release integrado do Hadoop. Mesmo hoje, alguns distribuidores oferecem versões customizadas, que nos lembram os anos iniciais do Linux. A curva de aprendizado íngreme é um dificultador, pois não se forma um técnico em Hadoop rapidamente e existem poucos especialistas no mercado.

Mas, tenho certeza de que em poucos anos ele será bem mais conhecido e utilizado. Uma pesquisa pelo termo Hadoop no Google Trends já aponta um crescimento significativo no interesse pela tecnologia, como podemos ver em http://www.google.com/trends?q=hadoop. Na minha opinião vale a pena investir tempo estudando a tecnologia. Um dos problemas com o Hadoop é realmente a sua complexidade de uso, demandando desenvolvedores altamente qualificados. A sua curva de aprendizado é bastante íngreme e praticamente inexiste gente habilitada para usar o Hadoop adequadamente. Bem, temos aí uma boa oportunidade para as universidades inserirem seu estudo e prática nos cursos de ciência da computação. Para um maior aprofundamento no Hadoop recomendo a leitura do livro “ Hadoop: The Definitive Guide”, publicado pela O’Reilly Media, de Tom White (White, 2010).

Mas como se desenvolve um projeto baseado em Hadoop? Começar um projeto Hadoop não é uma tarefa simples. Demanda a escolha de um caso que demonstre real valor para a empresa, necessita de profissionais capacitados e passa pela escolha de uma distribuição Hadoop adequada. Um projeto piloto que fracassa pode levar o conceito de Big Data ao descrédito e com isso atrasar a sua adoção na empresa.

O primeiro passo é identificar uma oportunidade que demonstre real valor para a empresa. Um caminho a seguir seria uma entrevista com executivos de negócio envolvidos em setores que poderão usufruir de Big Data, como áreas de risk management, detecção de fraudes, projetos de marketing que envolvam analise de sentimentos do cliente em relação à marca e aos produtos, análise da perda de clientes para concorrência e assim por diante. O projeto deve demonstrar o valor do conceito de Big Data e o papel importantíssimo do Hadoop. Com um resultado positivo, torna-se mais fácil estimular a disseminação do conceito de Big Data pela empresa.

Um segundo e importante passo é certificar-se de que existe capacitação adequada. Esta capacitação pode ser interna e/ou externa. Provavelmente, será necessário contratar expertise externa que transfira conhecimento para a equipe interna. O líder do projeto deve ser um profissional com ampla visão do negócio da empresa e bons conhecimentos de Hadoop. A equipe deve ter pelo menos profissionais com as seguintes características:

Função	Capacitação requerida
Administradores de sistemas	Cluster management, administração de Hadoop, análise de performance
Cientistas de dados	Estatística, programação SQL e Hive, expertise em garimpagem de dados e Business Intelligence, conhecimento da indústria e da empresa, facilidade de comunicação verbal e oral
Desenvolvedores	Java, Hive, Pig, MapReduce, Python e ferramental Hadoop que será usado no projeto

A terceira etapa é escolher a distribuição Hadoop. Como vimos anteriormente, existem diversas distribuições e é necessário escolher uma que tenha um ecossistema que garanta suporte e expertise para este projeto piloto. Lembre-se de que provavelmente não existe expertise interna na empresa e o distribuidor deverá suprir esta carência de expertise com recursos próprios (suporte e consultoria) ou com parceiros próximos.

Como Hadoop ainda está na infância, não existem muitos casos de sucesso que sirvam de benchmark. Portanto, é prudente não fazer investimentos pesados em infraestrutura, mesmo porque não existe muita experiência em especificação de infraestrutura para ambientes de

cluster em empresas comerciais. Este conhecimento está concentrado em universidades e centros de pesquisa, usuários típicos da computação de alto desempenho. Uma alternativa que sugiro é o uso da computação em nuvem para este projeto piloto. O cuidado a ser tomado é que Hadoop foi desenhado com a proposta de que é mais barato mover a computação que mover os dados. Assim, analise com cuidado o custo e o modo de transferir alguns terabytes para o provedor de nuvem. Os custos da computação em nuvem são baratos em termos de uso de servidores virtuais, mas podem encarecer muito quando movimenta-se grande quantidade de dados por redes de comunicação de e para o provedor da nuvem.

Dependendo do caso a ser usado como piloto, analise as fontes de dados necessárias e as formas obtê-las. É necessário definir um modelo de integração que envolva as diferentes origens e formas de dados, como obtê-las e como validar seu conteúdo.

E, finalmente, após o término do projeto, avalie os resultados. Será o modelo para futuros projetos de Big Data na empresa. Valide se realmente a escolha do caso foi adequada, se conseguiu-se demonstrar o real valor do conceito de Big Data, analise as carências de expertise e como preenchê-las para os próximos projetos, a adequação e escalabilidade da infraestrutura utilizada e assim por diante. O projeto piloto é um excelente estudo de caso para que Big Data passe a ser parte integrante das atividades da corporação.

Capítulo 8

Outras tecnologias de Big Data – Stream Processing e novas formas de visualização e animação de dados

Outra tecnologia que merece destaque em Big Data é o conceito de stream processing. Na verdade, se analisarmos o Hadoop e o Stream Processing vemos uma diferença fundamental. No Hadoop usamos processamento paralelo para analisarmos imensos volumes de dados que estão armazenados em centenas ou milhares de servidores. Já o stream processing analisa dados em movimento. O que o conceito de streams permite é que analisemos, por técnicas massivamente paralelas, imensos volumes de dados no momento em que estão sendo gerados, como uma corrente de dados que passa pelas regras de negócio de modo que possamos analisá-las, entendê-las em tempo real, para então definir ações imediatas em resposta.

A ideia de stream computing é fantástica! Um novo paradigma. No modelo de data mining tradicional, uma empresa filtra dados dos seus vários sistemas e após criar um data warehouse, dispara "queries". Na prática faz-se garimpagem em cima de dados estáticos, que não refletem o momento, mas sim o contexto de horas, dias ou mesmo semanas atrás. Com stream computing esta garimpagem é efetuada em tempo real. Em vez de disparar queries em cima de uma base de dados estática, coloca-se uma corrente contínua de dados (streaming data) atravessando um conjunto de queries. Podemos pensar em inúmeras aplicações, sejam estas em finanças, saúde e mesmo manufatura. Vamos ver este último exemplo: um projeto em desenvolvimento com uma

empresa de fabricação de semicondutores monitora em tempo real o processo de detecção e classificação de falhas.

Com stream computing as falhas nos chips sendo fabricados são detectadas em minutos e não horas ou mesmo semanas. Os wafers defeituosos podem ser reprocessados e, mais importante ainda, podem-se fazer ajustes em tempo real nos próprios processos de fabricação. Um outro exemplo simples seria o controle automático de tráfego, onde se pode coletar dados de sensores, de câmeras de vídeo, dados históricos e alertas da situação como acidentes ou vias em obras e, de acordo com um determinado algoritmo, identificar a probabilidade de um futuro congestionamento, atuando então de forma preditiva, alterando a temporização dos semáforos e desviando o fluxo de veículos para rotas alternativas. Na verdade é um caso real, desenvolvido pela IBM para a Autoridade de Trânsito de Singapura.

Este conceito é implementado por uma solução da IBM chamada de InfoSphere Streams. É um middleware ou uma infraestrutura de software, para desenvolvimento de aplicações que processam dados que são gerados continuamente ou ao vivo à medida que estes dados se tornem disponíveis (desde transações de "boca do caixa" a transações geradas por centrais telefônicas etc.).

A principal mudança conceitual proposta pelo InfoSphere Streams está na ideia de se processarem os dados à medida que eles são gerados, como transações de cartões de crédito criadas nos pontos de venda, com o intuito de, por exemplo, bloquear transações classificadas como fraudulentas em tempo real.

Naturalmente, a quantidade crescente de dados disponíveis em tempo real (streams de dados) vem aumentando rapidamente. No mercado financeiro de derivativos nos EUA, atingimos este ano a marca de um milhão de transações por segundo. Portanto, o modelo de construção de aplicações para processamento de dados desta natureza tem que evoluir de modo que se possa tanto acomodar a quantidade e a natureza destes dados, como a distribuição da carga de trabalho associada aos sofisticados algoritmos necessários para a análise e classificação destes dados.

Do ponto de vista dos clientes, a importância de aplicações desenvolvidas usando este paradigma é a rapidez e a capacidade de se antecipar a eventos de modo a diminuir custos (por exemplo, relativo a fraudes), tornar o negócio mais eficiente, diminuindo desperdício de materiais e energia (quando aplicado a processos manufatureiros), entre outros.

Claramente, novas oportunidades de negócios também são possíveis, devido à capacidade de processamento das informações em tempo real de dados disponíveis de sensores de

localização, como etiquetas de rádio-frequência (RFID), GPS disponível em telefones celulares, sensores em carros etc. Enfim, o potencial é quase inesgotável.

Concretamente, o InfoSphere Streams é um conjunto de componentes que gerenciam o ambiente de execução, incluindo desde a gerência de jobs (isto é, as aplicações que compartilham o ambiente), de gerência de recursos físicos (máquinas em um cluster), de gerência de segurança etc. Além disto, há também um ambiente de programação baseado em Eclipse (o Streams Studio), um ambiente de visualização de aplicações (StreamSight) e uma linguagem de programação chamada Stream Processing Language (SPL). Neste aspecto, o principal diferencial do Streams em relação a outros produtos é a flexibilidade desta linguagem, cujos conceitos básicos são os "operadores" (que fazem o papel de processar os dados que chegam, incrementalmente, por exemplo, fazendo um filtro ou classificando-os) e os "streams" que são sequências de tuplas ou registros, contendo as informações obtidas dos sensores que são manipuladas pelos operadores.

Portanto, tanto o ambiente de execução quanto a linguagem de programação são de uso geral. Esta abordagem torna possível a implementação de aplicações de diversas naturezas. Dentre os exemplos mais interessantes, podemos citar aplicações para o mercado de telecomunicações onde se colete call data records (CDR) e faça-se uma análise de padrões que permita predizer a chance de um grupo de pessoas migrar para competidores. Temos também exemplos de aplicações onde se monitoram os sinais biomédicos de recém-nascidos prematuros com o objetivo de se predizer se estão suscetíveis à infecção. É o exemplo do University of Ontario Institute of Technology (UOIT).

Quando se considera usar o InfoSphere Streams é importante que sejam bem identificadas as áreas onde existam a necessidade de se tornar proativo ao invés de reativo. De monitoração de ambientes manufatureiros, à análise de dados de tráfego de cargas e automóveis, à monitoração de redes de fornecimento de alimentos e farmacêuticos, à monitoração de redes de comunicação, transmissão de energia e de dados, dentre outros, sejam possíveis.

A principal vantagem competitiva que o sistema traz é a possibilidade de se processarem quantidades grandes de informações em tempo real, de modo incremental e "inteligente", usando a capacidade analítica do sistema. Para maiores informações sobre o InfoSphere Streams recomendo acessar o link http://www-01.ibm.com/software/data/infosphere/streams/.

O conceito de stream computing deve se disseminar devido às possibilidades que abre para aplicações em diversos setores, como detecção de fraude (impedindo uso mal-intencionado de cartões de crédito), monitoramento das redes de computação e de dados (cybersecurity), decifrando comentários e opiniões em mídias socias (sentiment analysis) e produtividade nas

linhas de produção das fábricas. Veremos nos próximos anos muitas empresas inovadoras surgindo neste setor. Um exemplo é a Guavus (http://www.guavus.com), que inclusive é usada por empresas de telecomunicações para identificar em tempo real como e por quem a rede está sendo usada.

Uma tecnologia que vai avançar muito nos próximos anos será a das técnicas de visualização e animação de dados. Visualização de dados tem como objetivo potencializar a apropriação da informação pelo usuário, por meio de recursos gráficos. A visualização de dados é uma área de aplicação de técnicas de computação gráfica interativas, que objetiva auxiliar a análise e a compreensão de um conjunto de dados.

Apresentar as informações de forma que os usuários possam consumi-la e extrair valor delas é fundamental para Big Data. A partir das visualizações e análises, que decisões e ações podem ser efetuadas. As técnicas de visualização vêm evoluindo muito e existe uma relação circular e intensa destas técnicas com o crescimento das demandas de análises de dados. Com maiores volumes, precisamos de novas formas de visualização de dados, que nos mostrem padrões antes irreconhecíveis e por sua vez estas novas técnicas de visualização incentivam o uso de mais análises. É um círculo virtuoso.

Entre as diversas técnicas temos as "tag clouds", que é a visualização em forma de lista visual ponderada, na qual as palavras que aparecem com maior frequência são visualizadas em caracteres maiores e destacadas. Este tipo de visualização permite facilmente identificar os termos ou palavras mais frequentes e portanto destacadas, em um determinado texto. Um bom exemplo de gerador de "tag clouds" é o wordle.net.

Outra técnica é o clustergram, usada para visualizar análises de cluster (cluster analysis) ou agrupamentos. Análise de agrupamentos é a classificação de objetos em diferentes grupos, cada um dos quais contendo os objetos semelhantes segundo alguma função de distância estatística. Esta classificação deve ser realizada de maneira automática, sem intervenção do usuário, sem considerar previamente propriedades características dos grupos e sem o uso de grupos de teste previamente conhecidos para direcionar a classificação. O link http://www.statajournal.com/sjpdf.html?articlenum=st0028 mostra a técnica em mais detalhes.

Outra técnica é a chamada "history flow", que mostra a evolução de um documento à medida que ele seja modificado por seus diversos colaboradores. O tempo aparece no eixo horizontal e as contribuições no eixo vertical. Cada autor tem uma cor diferente e o gráfico mostra quem colaborou e a grandeza desta colaboração, que é mostrada pelo tamanho da barra no eixo vertical. No link http://researcher.watson.ibm.com/researcher/view_project.php?id=3419 é possível vermos alguns exemplos bem interessantes de uso de history flow.

Temos também uma técnica chamada “spatial information flow” que visualiza de forma extremamente interessante e espacial informações específicas. Vejam o link http://senseable.mit.edu/nyte/ para exemplos inovadores desta técnica de visualização.

Visualização é um campo aberto para startups inovadoras. Diversas empresas surgem criando novas formas de visualizar dados e uma delas, a Zoomdata (www.zoomdata.com) cria visuais bem diferentes dos tradicionais visuais que as ferramentas de BI de hoje mostram para os seus usuários. Também apresenta uma interface muito mais intuitiva e fácil de usar. Na verdade, as ferramentas de BI atuais estão sob risco, caso se mantenham presas ao modelo de teclado e mouse da época de quando foram desenvolvidas. Os usuários estão em busca de novas formas de uso, que sejam fáceis e intuitivas para acessar e visualizar seus dados, como estão acostumados a fazer com seus smartphones e tablets. Existem diversas outras empresas focadas em visualização de dados como a Ayasdi (http://www.ayasdi.com/), Tableau (http://www.tableausoftware.com/) e Spotfire (http://spotfire.tibco.com/. Recomendo também assistir a dois vídeos do TED que abordam visualização de dados de forma inovadora: “A Beleza da visualização de dados” em http://www.ted.com/talks/david_mccandless_the_beauty_of_data_visualization.html, e “Visualizando nossa humanidade artisticamente” em http://www.ted.com/talks/aaron_koblin.html.

Um exemplo bem interessante de uso de visualização de dados é o Interactive American Migration Map, acessado em http://www.forbes.com/special-report/2011/migration.html.

Outro campo que deve avançar muito nos próximos anos é o uso de interfaces de voz, com o usuário perguntando e o computador respondendo em viva voz. Já existem casos bem-sucedidos como o Siri da Apple, que é um aplicativo que utiliza processamento de linguagem natural para responder perguntas, fazer recomendações e executar diversas ações. Na IBM, o Watson é um exemplo da futura interface de acesso a Big Data. Em vez de queries estruturadas, perguntamos em linguagem natural, como em uma conversação normal entre duas pessoas. Em 2011, a IBM apresentou o supercomputador Watson, promovendo uma competição de perguntas e respostas baseada no programa da TV americana chamado Jeopardy, **em que ele disputou** com os dois maiores vencedores desse jogo. O Watson foi concebido para entender o sentido da linguagem humana de acordo com o seu contexto, com o objetivo de encontrar a resposta precisa para perguntas complexas.

Sob esse ponto de vista, o programa Jeopardy oferece um grande desafio porque as perguntas não foram feitas para serem respondidas por um computador. Isso porque quem quiser participar desse programa precisa dominar todos os aspectos de uma linguagem natural, como regionalismos, gírias, metáforas, ambiguidades, sutilezas, trocadilhos e coisas do tipo e não

apenas trabalhar com o sentido literal da informação.

O Watson incorpora uma nova maneira de recuperar informação de forma rápida a partir de imensas quantidades de informação que lhe permitem ter uma profunda capacidade de análise e interpretação. De fato, a capacidade analítica de Watson é de investigar o equivalente a cerca de duzentos milhões de páginas de dados (ou perto de um milhão de livros) permitindo que ele seja capaz de responder a uma pergunta em aproximadamente três segundos. Essa capacidade de lidar com linguagem natural e responder precisamente questões complexas revela um grande potencial de transformar a maneira com que as máquinas interagem com os seres humanos, ajudando-os a conquistar seus objetivos.

Na próxima parte vamos analisar um aspecto fundamental para fazer Big Data gerar valor nas empresas: o perfil dos profissionais que lidarão com Big Data e a demanda por capacitação.

PARTE III

Recursos humanos para Big Data

Capítulo 9

Capacitação e perfis profissionais

O assunto Big Data começa a chamar a atenção. Diversas pesquisas apontam que muitas empresas começam a implementar iniciativas nesta área. Alguns estudos, como um recentemente efetuado pela Deloitte indica que essa tendência está apenas em estágio inicial de desenvolvimento e estima que menos de cinquenta grandes projetos (a partir de dez petabytes) estejam em execução em todo o mundo, quando da preparação deste livro, no primeiro trimestre de 2013.

Este cenário do crescimento do Big Data aponta também que estão surgindo novas oportunidades de emprego para profissionais de TI e de outros setores.

Um novo cargo, chamado de "data scientist" ou cientista de dados é um bom exemplo. Demanda normalmente formação em Ciência da Computação e Matemática, bem como as habilidades analíticas necessárias para encontrar a providencial agulha no palheiro de dados recolhidos pela empresa.

"Um cientista de dados é alguém que é curioso, que analisa os dados para detectar tendências", disse recentemente Anjul Bhambhri, vice-presidente de Produtos Big Data da IBM. "É quase como um indivíduo renascentista, que realmente quer aprender e trazer a mudança para uma organização."

Big Data abre oportunidades para a área de TI conseguir voar mais alto. Muitos CEOs expressam sua frustração com TI e uma frase de John Harris, chairman do Corporate IT Forum, organização que reúne altos executivos no Reino Unido, é muito interessante. Ele diz que há uma frustração muito grande entre os CEOs. Segundo ele, os CEOs sabem onde está o ouro e

não entendem porque TI não o extrai lá. Eles, CEOs, sentem que os gestores de TI não são geólogos que sabem onde extrair ouro. E faz uma comparação muito interessante com os técnicos que trabalharam na decifração da quebra dos códigos de comunicação dos alemães na Segunda Guerra Mundial. Eles eram matemáticos e linguistas que pensavam de forma criativa. Na sua opinião, os cientistas de dados devem ser os profissionais que conhecem profundamente o negócio e tenham imaginação e criatividade para fazer as perguntas certas. E não necessariamente serão encontrados no setor de TI.

Inédita há dois anos, a carreira de "cientista de dados" já aparece em profusão, pelo menos nos EUA. Como exemplo, acessei o http://www.Dice.com, um site americano especializado em carreiras de TI, em janeiro de 2013, ao escrever este capítulo e coloquei o termo "data scientist". Obtive 237 respostas.

O trabalho de um cientista de dados foi exemplificado na Harvard Business Review, versão online em outubro de 2012. No texto, os autores mencionam o trabalho de um pesquisador da Universidade de Stanford, que percebeu que a rede social LinkedIn estava monótona e que as pessoas realizavam poucas interações sociais. O pesquisador então sugeriu a criação de um algoritmo que apresentasse sugestões de amizades para os usuários da rede, também conhecido como People You May Know, o que foi um sucesso e ajudou com que a rede social se tornasse uma das mais utilizadas no mundo. O algoritmo proposto por Goldman utilizava as informações disponibilizadas nos perfis dos usuários da rede, como por exemplo, o colégio onde o usuário cursou o Ensino Médio. Comparando com os outros usuários, o algoritmo poderia sugerir pessoas que também estudaram no mesmo colégio, fazendo assim com que muitos aumentassem seu número de conexões, proporcionando maiores interações sociais pela rede. Este é um dos exemplos de como o cientista de dados utiliza as análises de dados do Big Data. O nome "data scientist", ou cientista de dados, foi utilizado pela primeira vez em 2008 e pode ser definido como um profissional de alto nível de formação, com curiosidade de fazer descobertas no mundo do Big Data. Uma outra definição pode ser aquele que utiliza dados e ciências para criar algo novo. Percebe-se uma grande tendência de aumento da necessidade destes profissionais no futuro, já que cada vez mais as empresas utilizarão as análises de dados como estratégia competitiva.

O cientista de dados vem emergindo como uma nova função, com escassez de profissionais e poucos cursos de formação. O cientista de dados vai trabalhar em uma disciplina que podemos chamar de "Data Science" ou "Ciência dos Dados". Este é o grande desafio do Big Data nos próximos anos. Ter profissionais capacitados, uma vez que a tecnologia está evoluindo rápido e não será impeditiva. O gargalo não é tecnologia, mas gente. À medida que Big Data se insere nas empresas, os próprios conceitos de gestão, baseados em "orientação a suposições" passarão a ser orientados a fatos. A razão é simples: um imenso volume de dados permitirá

fazer análises antes inimagináveis sobre dados, examinando fatos e fazendo previsões com muito mais precisão. Estas análises preditivas demandam uma capacitação que envolve estatística, matemática e conhecimento de negócios, que é bem diferente das atividades dos analistas envolvidos com ferramentas de BI hoje, que estão mais preocupados em criar gráficos e dashboards para mostrar dados passados. Hoje a maioria das ações de BI envolve dados armazenados em data warehouse ao longo do tempo e apenas consegue visualizar retrospectivas. Chegar a análises preditivas é um passo que não se dá de um dia para o outro.

Como é uma função nova, claro que surgem definições pouco claras e profissionais que sabem usar ferramentas de BI começam a se autointitular data scientists. Para chegar a serem cientistas de dados precisam demonstrar capacitação adequada para isso e não apenas o conhecimento de ferramentas de BI. Uma comparação de skills mostra a diferença. Um profissional de BI geralmente tem capacitação em ferramentas como Cognos, data warehouse, uso de SQL e conhecimentos de bancos de dados relacionais, como SQLServer, Oracle ou DB2. O cientista de dados precisa ter conhecimentos de estatística, matemática, entender do negócio e ter familiaridade com tecnologias e linguagens como Hadoop e Pig. Para os profissionais engajados com Big Data aparece um novo desafio que é a modelagem de dados não estruturados. Nos últimos trinta anos, os arquitetos envolvidos com modelagem de dados se especializaram no modelo relacional, suas regras e técnicas. Por exemplo, temos eliminação de redundâncias através da normalização como também critérios rígidos de garantia de integridade referencial. Bancos de dados NoSQL não se preocupam com duplicação de dados e não exigem regras de integridade referencial.

Podemos ver as diferenças entre os perfis profissionais no quadro adiante, que pode ser usado como roteiro para um analista de BI se tornar um cientista de dados:

Analista de BI	Cientista de dados
Cognos, modelo relacional, banco de dados SQLServr, Oracle, DB2	Hadoop, modelos relacionais e NoSQL, bancos de dados não relacionais e in-memory
Modelagem relacional/estruturada	Inclui também modelagem não estruturada. Modelagem analítica é essencial.
Desenvolve queries estruturados sobre dados passados.	Cria perguntas e busca relacionamentos entre fatos aparentemente desconexos.

Mas, além do data scientist, existe espaço para outras atividades profissionais. Por exemplo, haverá forte demanda também por desenvolvedores e administradores de sistemas que se especializam em ferramentas voltadas para Big Data, como o Hadoop, tecnologia projetada para aplicações distribuídas com uso intensivo de dados e utilizadas por sites bastante conhecidos como o Yahoo, Facebook, LinkedIn e eBay. O Hadoop também já é mencionado em muitos dos anúncios dos empregos disponibilizados na Dice.com. Coloquei o termo Hadoop em janeiro de 2013 e recebi 922 respostas de cargos como engenheiro ou desenvolvedor de software com este requisito.

Em resumo, podemos identificar três perfis básicos de profissionais engajados em Big Data:

a. Cientistas de dados, como descrevemos antes. Profissionais capacitados em estatística, ciência da computação e/ou matemática capazes de analisar grandes volumes de dados e extrair deles insights que criem novas oportunidades de negócio;
b. Analistas de negócio que, conhecendo bem o negócio em que atuam, consigam formular as perguntas corretas. Analisar as respostas e tomar decisões estratégicas e táticas que alavanquem novos negócios ou aumentem a lucratividade da empresa. Esta função tende a ser acoplada às funções do cientista de dados.
c. Profissionais de tecnologia que cuidarão da infraestrutura e seu suporte técnico para suportar Big Data. O aparato tecnológico de Big Data não é muito comum em empresas tipicamente comerciais, pois demanda expertise em gerenciar hardware em clusters de alta performance (Hadoop é massivamente paralelo) e pensar em volumes de dados significativamente maiores e muito mais variados que comumente se usam em sistemas tradicionais.

Entretanto, nos próximos anos viveremos uma escassez destes profissionais, não só no Brasil, mas no mundo todo. Esta escassez, ao mesmo tempo que abre muitas perspectivas profissionais para os que abraçarem a função, também atuará como um entrave, pois dificultará às empresas usarem Big Data com eficiência. Recentes pesquisas estimam que, por volta de 2015, Big Data demandará cerca de 4,4 milhões de profissionais em todo o mundo e que apenas 1/3 destes cargos poderá ser preenchido com as capacitações disponíveis hoje em dia. Uma pesquisa mundial da IBM corrobora estes dados, mostrando que apenas uma em dez organizações acredita que tenha profissionais com as capacitações necessárias e que três em cada quatro estudantes e professores reportam que existe um gap de moderado a grande entre o que é ensinado hoje e o que o mercado de trabalho realmente necessita.

Portanto, os principais desafios que as empresas terão serão relacionados a como recrutar e desenvolver estes profissionais.

Algumas já usam conceitos de Big Data em diversas atividades, como bancos e empresas de cartão de crédito em seus sistemas de detecção e combate a fraudes. Big Data será uma expansão destes conceitos. Entretanto, em muitas outras, poucas ações de uso analítico de dados acontece hoje, e geralmente, quando acontece, é em projetos de setores bem específicos, como em determinadas ações de marketing. Para estas empresas, Big Data será um mundo novo, demandando um perfil profissional que inexiste hoje em seus quadros.

A academia tem um papel de fundamental importância. A formação sólida de profissionais não pode ficar exclusivamente a cargo dos fornecedores de tecnologias pois estes tendem a focar a capacitação no uso de suas tecnologias e não nos conceitos fundamentais. A capacitação adequada em estatística, matemática e análises avançadas de dados deve ser feita pela academia e complementada na prática pelos fornecedores através do uso de suas tecnologias. A educação em Big Data não deve ser restrita apenas aos profissionais diretamente envolvidos, mas deve abranger toda a organização. Os executivos C-level devem ter compreensão do potencial de Big Data para usá-lo em seus processos rotineiros de decisão. Os executivos do escalão médio devem entender que Big Data pode ajudá-los a identificar problemas e mesmo redesenhar os processos de negócio pelos quais são responsáveis. E os profissionais de TI têm que entender que, mesmo sem estarem diretamente envolvidos com Big Data, estarão em constante intercessão com o conceito. Big Data vai ser parte integrante da arquitetura de dados e aplicações das empresas.

Um exemplo de atuação complementar entre academia e fornecedor de tecnologia é a da IBM, que tem uma nova iniciativa nos EUA, denominada Big Data University, que visa a formação de estudantes de graduação e pós-graduação na área, expondo-os ao Hadoop e aos conceitos de Big Data. Lançada em outubro passado, a Big Data University já atraiu mais de 18 mil estudantes para seus cursos online gratuitos, em inglês. Deem uma olhada em http://www.bigdatauniversity.com/.

Capítulo 10

O papel do CDO (Chief Data Officer)

A função de gestor do Big Data pode ser denominada Chief Data Officer (CDO) e é uma posição praticamente desconhecida na maioria das empresas. Deve ser uma posição de nível senior e, como as demais posições denominadas C-level, deverá estar em constante evolução.

O quadro mostra a função CDO em comparação a alguns cargos C-level:

Posição	Origem	Função atual/ futura	Quando surgiu
CFO (Chief Financial Officer)	Gerentes de contabilidade, controllers	Foco na criação de valor para os acionistas	Início dos anos 70
CIO (Chief Information Officer)	Gerentes de CPD (centros de processamento de dados)	Impulsionar transformação dos negócios através de TI	Meados dos anos 80
CISO (Chief Information Security Officer)	Responsáveis pela segurança física, proteção de dados e compliance	Gerenciar riscos e segurança de modo a otimizar as operações do negócio	Meados dos anos 90

CDO (Chief Data Officer)	Analistas financeiros, analistas de Business Intelligence	Impulsionador de valor para a organização através da análise avaxnçada de dados, gerando vantagens competitivas	Está surgindo agora. Deve se disseminar em um horizonte de pelo menos cinco anos

Apesar da importância do conceito de Big Data, a criação desta função em nível senior encontra algumas barreiras organizacionais, como a ainda falta de compreensão do impacto estratégico e do valor de Big Data para obter vantagem competitiva, a falta de capacitação para exercer tal função no quadro gerencial da empresa e no mercado, questões culturais impeditivas ao incentivo do compartilhamento de dados e a carência de suporte executivo. Muitos CEOs (Chief Executive Officers) ou presidentes/gerentes gerais das empresas ainda não têm a plena percepção do valor do Big Data e tendem em muitas situações a encarar como mais uma tecnologia e portanto afeita à área de tecnologia da organização. O CDO não deve ficar subordinado ao CIO, mas atuar em parceria com ele e em estreito relacionamento com os executivos C-level da empresa. É o vínculo de ligação entre TI e os negócios, sob a forma de geração de valor explorando dados internos e externos. Além disso, é importante frisar que sua função é muito mais centrada no negócio que na tecnologia.

Algumas de suas principais responsabilidades são:

a. Atuar de forma integrada com os demais executivos C-level para determinar necessidades estratégicas da empresa e desenhar ações de estratégia analítica para atender a estas demandas. Análises avançadas não funcionam em silos isolados. Apenas quando as áreas de negócio trabalham em conjunto é que insights relevantes são gerados.
b. Orientação no uso das ferramentas de análise. Atuar como evangelizador do conceito de Big Data e criar um road map para a estratégia de exploração dos dados na empresa. Desenhar programas de capacitação e educação em analítica avançada.
c. Criar e gerenciar o Centro de Excelência em Analítica Avançada, onde desenvolverá as ações de evangelização, capacitação, disseminação e suporte no uso de analítica avançada.
d. Ajudar a organização no redesenho e refinamento de processos baseados no uso de Big Data como, por exemplo, análises em tempo real influenciando os processos durante a execução destes.

A função CDO deve estar como os demais C-level, subordinadas diretamente ao CEO da organização. À medida que a organização amadureça no uso de Big Data, a função passa a ter métricas que avaliem o seu resultado para o próprio negócio. Na minha opinião estas métricas,

obviamente refinadas ao longo do tempo, devem começar com a própria criação da função.

Para apresentar uma ideia mais prática do que é a atuação do CDO, quero compartilhar aqui a conversa que tive, em meados de 2012, com Mario Faria, o primeiro CDO (Chief Data Officer) do país, pois creio ser extremamente válido expor a experiência dele nesta posição pioneira.

1. Mario, como primeiro CDO (Chief Data Officer) no Brasil você está abrindo novos caminhos. Existe muita curiosidade sobre o assunto e creio que podemos conversar um pouco sobre o tema. Antes de mais nada o que é exatamente um CDO, quais suas funções e responsabilidades e onde ele deve se posicionar na organização?

Mario: Minha função é bastante nova, apesar das empresas se preocuparem com o assunto dados há décadas. O papel de um CDO é ser o responsável por gerir os dados da empresa, através de uma estratégia baseada em valor para o negócio. Mesmo nos Estados Unidos, esta posição é nova, e o primeiro CDO foi o Professor Richard Wang do MIT, que em 2010, se licenciou para ser o CDO do Exército Americano.

O meu papel é conseguir olhar para as necessidades que a empresa tem em desenvolver novos produtos, serviços e ofertas, e quais são os insumos (no caso os dados) que precisam estar disponíveis para que isto ocorra. Se eu trabalhasse em uma indústria, meu cargo seria o de Diretor de Materiais.

2. Quais são as características e skills necessários a um CDO? Existe alguma educação formal?

Mario: Antes de mais nada, um CDO precisa gostar de gente, pois vai ter que conversar e interagir muito com as áreas de negócio da empresa, a equipe de tecnologia e os principais clientes. Depois, o CDO tem que conhecer tecnologia e estar antenado nas grandes tendências do setor. E para finalizar, ter um raciocínio lógico e conhecer bastante de processos.

Eu tenho formação em Computação, mestrado também em Computação e um MBA em Marketing. Tudo isto ajuda, porém não é suficiente para um excelente resultado.

3. O CDO substitui ou complementa outras funções como analistas de negócios ?

Mario: No meu caso, por estar me reportando diretamente ao CEO da empresa, sendo par do CIO e dos principais executivos de Vendas, Produtos e Operações, vejo minha função como parte integrante do sucesso da empresa. Os analistas de negócios têm uma função muito

específica que é traduzir as necessidades em uma linguagem que o pessoal de TI consiga implementar. Talvez o CDO e sua equipe sejam os analistas de dados, para fazer uma analogia

4. Porque você entrou nesta linha de atuação profissional?

Mario: Foi um convite feito pelo CEO da empresa, Dorival Dourado, para ajudá-lo a construir uma empresa séria, focada e de sucesso que é a Boa Vista Serviços. A minha posição existe pela interação que a Boa Vista teve com o Professor Richard Wang em 2011, onde ele recomendou que pelo nosso negócio, deveríamos ter uma área específica focada em dados, com um executivo dedicado a este assunto. Como eu adoro startups e desafios, abracei a oferta na hora.

5. Na sua opinião quais os desafios que o CDO enfrenta em uma empresa e como você sugere resolvê-los?

Mario: O maior desafio é mostrar o que um CDO e uma área de dados faz, e quais benefícios poderá trazer. Depois é entender que, apesar de existir um componente de tecnologia, esta função não é TI e possui um foco bem definido em olhar e tratar o ciclo de vida do dado na empresa. E o mais importante, o CDO é um prestador de serviços para Vendas, Marketing, Produtos e Operações, e deve estar sempre atento como pode estar a um passo adiante das necessidades destas áreas.

6. Que recomendações você faria a quem quer se tornar também um CDO?

Mario: Vou compartilhar o que eu fiz quando entrei aqui. Estudei muito, li bastante, pesquisei sobre este assunto e falei com o máximo de pessoas que consegui. Foi uma dedicação intensa. Além disto, dediquei um tempo significativo para entender a empresa e as áreas com as quais eu iria me relacionar e interagir. E acima de tudo, tenho sido aberto a escutar a minha equipe, que tem ajudado bastante nesta tarefa de construir nossa área, que é nova e cheia de desafios.

7. Quais tipos de empresas deveriam também ter um CDO ?

Mario: O fato que é hoje vivemos um momento peculiar da história humana, onde a quantidade de dados gerados é infinitamente superior ao que é consumido. Estes dados, em sua maioria não são estruturados, e são criados em uma velocidade tremenda. Isto é o conceito que o mercado batizou de Big Data. Toda empresa que está olhando para isto, como um fator disruptivo na sua indústria, precisa ter um CDO. Não vejo como bancos, varejo, empresas de telecom e empresas que utilizam a Internet como um meio para fazer negócios conseguirão ficar mais que dois anos, a partir de hoje, sem um CDO. Segundo o IDC, o mercado de Big Data irá movimentar quase

dezesseis bilhões de dólares em 2015. É um oceano de oportunidades para todos.

Como vimos na conversa com Mario, este é um dos grandes desafios do Big Data: a escassez de profissionais qualificados para tratar analiticamente as informações geradas por estas imensas bases de dados. Um artigo interessante, publicado pelo Wall Street Journal, edição brasileira, destaca claramente este problema: "MBAs agora preparam mineiros de dados", podendo ser acessado em http://online.wsj.com/article/SB10001424053111903480904576510934018741532.html.

O artigo diz que, diante do fluxo crescente de dados da Internet e outras fontes eletrônicas, muitas empresas começaram a procurar gerentes que saibam interpretar os números usando uma prática em expansão: a análise de dados, também conhecida como inteligência empresarial. Mas, encontrar profissionais qualificados tem-se mostrado difícil. Daí que, nos próximos meses, várias faculdades americanas, como a Faculdade de Pós-Graduação em Administração da Universidade Fordham e a Faculdade de Administração Kelley, da Universidade de Indiana, começam a oferecer disciplinas eletivas, cursos de extensão e mestrados em análise de dados; outros cursos e programas do tipo foram lançados no ano passado.

A análise de dados já foi considerada tarefa de especialistas em matemática, ciência e tecnologia da informação (TI); mas, diante da enxurrada de dados da Internet e outras fontes, as empresas demandam agora profissionais capazes tanto de analisar informações, como também de ajudar as empresas a resolver problemas e criar estratégias. Com certeza, é uma boa oportunidade profissional.

Conclusões

Capítulo 11

Comentários Finais

Vencer os desafios do Big Data é essencial para as empresas se manterem competitivas na economia digital. Estamos ainda no início da sua curva de aprendizado, mas é fundamental que as ações comecem de imediato. A velocidade com que as mudanças nas tecnologias e no cenário de negócios acontecem não nos permite o luxo de ficar esperando para ver o que virá.

Big Data nos abre o que podemos chamar de portas para uma "intelligent economy" ou economia inteligente que produz um fluxo contínuo de informações, que podem ser monitoradas e analisadas. Dados do instituto de pesquisas IDC nos mostram que, em 2012, o volume de informações criadas (e replicadas) ultrapassou a casa dos 2,7 zetabytes e deve chegar a oito zetabytes em 2015. Estas informações são geradas por diversas fontes, que vão dos tradicionais desktops, passando por tablets e smartphones, mídias sociais e os milhões de sensores que se espalham pelas infraestruturas das cidades e da sociedade.

Com Big Data as empresas podem usar dados transacionais e não-transacionais para traçar estratégicas, decisões comerciais de longo prazo sobre, por exemplo, o que e quando colocar nas prateleiras das lojas. Big Data tem papel importante na economia de um país, pois torna suas empresas mais competitivas. Assim, o governo poderia colaborar com incentivos para a capacitação e criação de startups voltadas para Big Data, pois informação passa a ter papel cada vez mais estratégico não só para empresas como para nações. O governo também deveria enfatizar uma política de abertura de dados públicos, criando mecanismos de integração entre as suas diversas bases de dados, o que possibilitaria mais transparência e uso mais inteligente destes dados para tomada de decisões, não só governamentais, como empresariais. Muitos dados públicos são de grande importância para a tomada de decisões corporativas.

Importante lembrar que Big Data não acaba com Business Intelligence (BI), mas o torna mais valioso e útil para o negócio. Na prática, sempre teremos a necessidade de olhar para o

passado e com a possibilidade de analisar um grande volume de dados, BI vai ser reforçado.

As decisões de negócio não serão tomadas exclusivamente com base na análise do Twitter e Facebook, por exemplo, mas usando-se todas as fontes de informação. Mas as mídias sociais podem gerar insights muito interessantes. Por exemplo, se você aprender pelos tuítes e likes do Facebook como os fãs de determinado artista se comportam como consumidores, comprando blusões ou calças iguais aos que ele se apresenta nos shows, você poderia, então, tomar a decisão de estocar esses produtos apenas nas cidades dos futuros shows, uma vez que será uma mercadoria de venda por tempo muito limitado. Assim, uma empresa pode aproveitar a janela de oportunidade com uma análise preditiva do mercado, procurando padrões nos tuítes e likes que se correlacionam com marca e localização.

Esta é a diferença em relação ao BI tradicional, onde as decisões têm como base apenas dados históricos e, portanto, impossíveis de prover momentos instantâneos como esses.

As tecnologias de Big Data permitem que a informação seja trabalhada antes de ser otimizada, racionalizada ou relacionada. Isso, juntamente com análise avançada, permite fazer e responder perguntas em ciclos muito curtos. Com as tecnologias de Big Data podemos gerar milhares de modelos de dados para cada linha de produto, com previsões de vários meses à frente. Não muito tempo atrás, isso seria quase impossível. Levaria semanas ou até mesmo meses para que analistas estatísticos construíssem um único modelo.

Estes cenários mostram por que o mercado de Big Data está cada vez mais aquecido. Algumas estimativas indicam que ele crescerá a uma média de 40% ao ano nos próximos três a quatro anos. O IDC acredita que em 2020, quando a indústria de TI global como um todo alcançar a cifra dos cinco trilhões de dólares, ou seja, 1,7 trilhões maior que hoje, cerca de 80% do crescimento desta indústria se dará em torno das novas ondas tecnológicas convergentes como mobilidade, cloud computing, social business e claro, Big Data.

Este interesse se reflete em números muito interessantes, como o cerca de meio bilhão de dólares investido por venture capitalists em startups de Big Data nos últimos anos. Por outro lado, a tendência é que o mercado de tecnologias de Big Data fique concentrado nos grandes atores já existentes hoje na indústria de software, como IBM, HP, Oracle e Microsoft. Uma simples olhada no mercado mostra as grandes empresas adquirindo empresas inovadoras nesta área, como a IBM comprando Algorithmics, Demandtec, Emptoris, Varicent, Vivisimo e outras, HP comprando Vertica e Autonomy, e Oracle adquirindo Endeca. Isto apenas em um curto espaço de tempo!

Uma variável importante que ajuda a delinear este cenário é que a maior parte dos

investidores em startups de Big Data tem optado, como estratégia de saída, pela venda das empresas investidas a um destes grandes atores. Claro, sempre existirá espaço para empresas menores, de nichos especializados. O cenário de Big Data, pelo menos durante esta década, será recheado de inovações e startups criando novos tipos de aplicações. Os investidores continuarão aportando investimentos em startups neste setor e sinalizando grandes oportunidades para empreendedores criativos.

Uma área que demandará crescimento significativo será em serviços de consultoria, devido à escassez de capacitação em análises avançadas. A falta de capacitação será um fator inibidor para o crescimento do mercado de Big Data e isso abre grandes oportunidades para consultorias especializadas. O mercado de consultoria será preenchido pelas grandes consultorias e integradores, bem como empresas "butique", especializadas em determinados nichos de mercado.

Big Data abre um novo cenário de oportunidades para as empresas, permitindo que elas analisem dados, em volume e variedade de formas inimagináveis há alguns anos atrás, criando vantagens competitivas significativas. Big Data não é em absoluto um hype de mercado. É um tsunami ainda em alto mar, pouco visível, mas com poder de causar devastação imensa se for ignorado. A sugestão que faço é avaliar o impacto do Big Data na sua indústria e sua empresa e quão distante a sua organização está hoje em termos de "estar preparada" para o que vem pela frente. Isto significa avaliar a empresa e a área de TI para as tecnologias, capacitações e processos que serão necessários para explorar o potencial do Big Data.

Adicionalmente ainda é um cenário imaturo, havendo poucos exemplos de "melhores práticas". Portanto, é uma iniciativa inovadora para a maioria das empresas, com os riscos e claro, as recompensas dos empreendedores inovadores. Mas, ficar parado esperando a onda chegar será perigoso, pois, provavelmente até o fim da década Big Data passará a ser apenas "Just Data". Será o modelo natural de pensar análises de dados. Neste momento, Big Data se tornará ubiquo nas empresas e o termo Big deixará de fazer sentido. Os dados existirão naturalmente em volumes significativos, mas os outros V (velocidade, variedade, veracidade e valor) estarão combinados para gerar novos processos e novas maneiras de encarar este imenso potencial das empresas que são os dados. Tratar e analisar dados será tão importante para as organizações quanto os demais fatores, como recursos humanos, tecnológicos e financeiros. As empresas simplesmente não viverão sem analisar dados continuamente.

Bibliografia

Como vocês observaram, inseri no texto muitos links que ajudarão no aprofundamento em algum tópico que desperte mais interesse. Estes links levam a artigos e blogs que descrevem os temas com mais profundidade. Recomendo que sejam acessados, pois levam à informações complementares bem relevantes. O uso intenso de links complementares ajuda a diminuir o tamanho do livro e permite uma leitura mais rápida, possibilitando que o leitor se aprofunde apenas nos temas que tiver mais interesse.

Além disso, sugiro a leitura de algumas publicações que complementam o material deste livro. Recomendo o livro, cujo download é gratuito, chamado "Understanding Big Data", escrito por cinco profissionais senior da IBM. Pode ser acessado em https://www14.software.ibm.com/webapp/iwm/web/signup.do?source=sw-infomgt&S_PKG=500016891&S_CMP=is_bdebook1_bdmicrornav.

A IBM mantém um portal exclusivo para debater Big Data, com vários papers que podem ser baixados gratuitamente. O acesso é http://www-01.ibm.com/software/data/bigdata/.

Recomendo especificamente a leitura de um relatório "Analytics: the real-world use of Big Data" em http://www-935.ibm.com/services/us/gbs/thoughtleadership/ibv-big-data-at-work.html.

Recomendo também o relatório produzido pela McKinsey Global Institute (MGI), "Big Data: The next frontier for innovation, competition, and productivity", disponível em http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovat O download também é gratuito.

Outra leitura recomendada é o relatório do World Economic Forum de 2012, "Big Data, Big Impact: new possibilities for International Development", que pode ser baixado em http://www3.weforum.org/docs/WEF_TC_MFS_BigDataBigImpact_Briefing_2012.pdf.

Um paper sobre visualização que sugiro ler é "Harness the Power of data visualization to transform your Business" acessado em http://www.tentonmarketing.com/Portfolio/SAS-Data-

Visualization_WP.pdf.

Existem também bons livros sobre o assunto, dos quais destaco aqui:

"Big Data Analytics: disruptive technologies for changing the game", de Dr. Arvind Sathi. Baixei no meu Kindle e o li bem rápido, pois são menos de cem páginas.

Outros que li (todos pelo Kindle) e cuja leitura recomendo são:

"Predictive Analytics", de Eric Siegel;

"Big Data: a revolution that will transform how we live, work and think" de Viktor Mayer-Schonberger e Kenneth Cukier;

"Privacy and Big Data" de Terence Craig e Mary Ludloff.

Algumas sugestões adicionais:

Blog sobre Big Data no Forrester Research: http://blogs.forrester.com/category/bigdata;

Um exemplo de uso na medicina em http://www.slideshare.net/sigindia/fs-big-science-big-data-big-collaboration.

Na mídia de negócios e tecnologia encontramos muitos artigos interessantes. Como sugestão recomendo acessar os sites da Businessweek, New York Times, Forbes, CIO.com, Computerworld, Economist, Walls Street Journal, Information Week e na caixa de search coloquem "Big Data". Vão aparecer artigos muito bons.

Muitos fornecedores de tecnologia como EMC, Oracle, Microsoft e HP e consultorias como Deloitte, Booz Allen e Accenture fornecem soluções para Big Data e em seus sites disponibilizam muitos artigos interessantes. Uma busca neles por Big Data vai retornar textos que valem a pena serem lidos. E claro, uma busca no Google, usando o termo Big Data vai retornar muita coisa. Em final de março retornou cerca de 947 milhões de respostas. Claro que nem tudo vale a pena, mas um refinamento na busca filtra melhor o conteúdo. Uma sugestão de como refinar as buscas pode ser vista neste texto em http://searchengineland.com/guide/how-to-use-google-to-search.

Recomendo também acessar alguns vídeos do TED, como:

“Hans Rosling mostra as melhores estatísticas que você já viu” em http://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen.html;

“A Beleza da visualização de dados” em http://www.ted.com/talks/david_mccandless_the_beauty_of_data_visualization.html;

“Visualizando nossa humanidade artisticamente” em http://www.ted.com/talks/aaron_koblin.html.

Enfim, artigos, livros e papers sobre Big Data não faltam. O desafio é conseguirmos tempo para lê-los...