BIG DATA E GESTÃO

Bernardo F. E. Lins[1]

1. Big data: uma realidade em nossas vidas

Big Data é uma expressão que se refere a um conjunto de técnicas para a coleta e tratamento de grandes volumes de dados, com um objetivo de análise preditiva. Trata-se não apenas de operar sobre repósitórios volumosos, com grande variedade de informações e formatos, um aspecto que é facilmente compreendido por todos. Trata-se, também de fazê-lo com uma mentalidade exploratória e com uma flexibilidade de tratamento que distanciam essas técnicas da estruturação convencional de sistemas de gerenciamento de bancos de dados, uma abordagem que em geral soa um pouco hermética.

A estruturação dos dados, sua coleta e seu armazenamento devem, evidentemente, existir no substrato do *Big Data*. No entanto, é preciso deixar claro que as ferramentas de *Big Data* não estão voltadas para dar a esses dados o tratamento para o qual foram coletados. O enfoque é justamente o oposto: coletar, construir e cruzar dados para obter informações ou *insights* para os quais esses dados **não** foram originalmente concebidos.

Por essa razão, as técnicas correlatas a *Big Data* representam um desafio. Fala-se, nesse contexto, de dados que são intensamente manipulados e cruzados, permitindo a construção de correlações muitas vezes inesperadas e revelando aspectos surpreendentes do mercado, das pessoas e das políticas. Ademais, em geral esses dados são perecíveis, são continuamente coletados e, em muitos casos, não permanecem armazenados.

Para examinar, em uma primeira aproximação ao problema, o uso desses dados, o modo como são aproveitados e as implicações desses aspectos para a empresa, tanto em termos de tecnologia quanto em termos de legislação, este texto está organizado como se descreve a seguir. Na seção 2, apresentam-se alguns elementos descritivos do que seja *Big Data* e das principais abordagens que alcança. Na seção 3, discute-se a estratégia para preparar-se para trabalhar com *Big Data* e destacam-se quatro exemplos em que a técnica teve sucesso: os serviços de catálogo na internet, a investigação criminal, a orientação de campanhas eleitorais e o acompanhamento de políticas públicas de saúde. Na seção 4, faz-se um breve comentário sobre os custos envolvidos na adoção dessa técnica. Nas seção 5, é esboçado um quadro sobre o uso dessas técnicas na gestão global de um empreendimento e o problema da proteção de dados pessoais é formulado. Apresentam-se, enfim, as conclusões.

---

[1] Engenheiro civil e doutor em economia. Membro acadêmico da ABQ. O autor agradece as contribuições a este ensaio de José de Souza Paz Filho, mestre em engenharia elétrica e consultor legislativo da Câmara dos Deputados e de Jerônimo Cabral Guedes, engenheiro eletricista e consultor empresarial,

2. *Big Data*: afinal, o que é isso?

O problema de colher, armazenar, organizar e trabalhar com dados acompanha a computação desde suas origens. Mesmo as aplicações mais triviais da informática envolviam alguma forma de tratamento de dados: textos, registros de contabilidade, informações de pagamentos, medidas de peças torneadas em um equipamento, dados de identificação de clientes, distâncias percorridas por veículos e inúmeros outros exemplos, de acordo com cada aplicação processada em computadores. Isto levou a um contínuo desenvolvimento de métodos de arquivamento e acesso em diferentes mídias (cartões, fitas perfuradas, fitas magnéticas, discos, disquetes, HD e assim por diante). Eventualmente, a partir da década de 1960 foram desenvolvidos sistemas de gerenciamento de bancos de dados (SGBD) que permitiam não apenas armazenar, mas também relacionar os dados arquivados, viabilizando formas mais eficazes de organização e recuperação de informações (GODA e KITSUREGAWA, 2012).

A noção de mineração de dados como abordagem exploratória foi introduzida ao final da década de 1980, a partir da constatação de que as empresas que faziam uso da informática acumulavam grandes estoques de dados que permaneciam em repouso, sem uso, às vezes por longos períodos. Esses dados teriam potencial para subsidiar análises de mercado e processos de engenharia ou marketing, embora em muitos casos não houvessem sido coletados com este fim. Por exemplo, dados da história contábil da empresa poderiam ajudar a revelar a produtividade e desempenho de seus processos operacionais e de negócio; no entanto, não haviam sido pensados para isso e efetivamente não eram usados para isso. A mineração de dados é a descoberta de padrões novos ou correlações inesperadas em grandes conjuntos de dados, que podem ser usados para resolver novos problemas (McCUE, 2015: 31).

O fundamento da mineração de dados, além da pesquisa em diferentes bases que originalmente não se correlacionavam, envolvia uma inversão na lógica de trabalho do profissional de informática. Em lugar de se desenvolver aplicações, organizar e colher dados para atender a requisitos ou exigências do negócio estabelecidas *ex-ante*, fazia-se uma pergunta de pesquisa capaz de produzir inovação no negócio. A partir dessa pergunta, examinava-se que tipo de informação era necessária para que esta fosse respondida e onde esses dados poderiam ser recuperados. A abordagem envolveria, então, uma descoberta de dados que pudessem ser úteis.

Um episódio relatado por Mayer-Schönberger e Cukier (2013: 2-3) ilustra essa diferença. Por volta de 2009, ocorreu o primeiro surto de H1N1 a partir do México e esperou-se, nos meses seguintes, uma pandemia. Nos EUA, os centros para controle e prevenção de doenças (CDC) mantiveram, nas semanas seguintes, um acompanhamento dos registros médicos de ocorrências da gripe, com a expectativa de identificar eventuais surtos. Os dados eram tabulados semanalmente e era possível acompanhar a evolução geográfica da doença, ainda que com uma quinzena de atraso.

Paralelamente, engenheiros da empresa Google procuraram uma abordagem alternativa. Examinaram cerca de 50 milhões de termos de pesquisa mais comumente usados pelos usuários e os compararam com os mapas de evolução de doenças do CDC nos cinco anos anteriores. Um sistema desenvolvido por eles correlacionou essas expressões de busca, sem levar em conta seu significado, com as ocorrências de doenças na

mesma localidade do usuário. Após experimentar com dezenas de milhões de diferentes modelos de regressão, chegaram a um conjunto de 45 termos que se correlacionavam com a ocorrência de doenças. Desse modo, acompanhando a atividade dos usuários dali em diante, o Google foi capaz de apontar, na medida em que as pesquisas eram submetidas ao portal, que localidades vinham sendo alcançadas pela gripe. Com uma diferença: a identificação era imediata, não havia atraso.

A abordagem dos CDC era convencional. A do Google, era de mineração de dados. E esta última era muito mais eficiente em termos de resultados, sendo esse o ponto positivo. No entanto, era evidentemente muito mais cara de se desenvolver, e esse é o ponto fraco.

A forma de trabalhar ilustrada no exemplo não exaure as possibilidades da mineração de dados. Um passo ulterior seria dado ao se dispor de ferramentas cada vez mais avançadas e processadores mais rápidos. Em lugar de se ter uma pergunta como ponto de partida, o analista dispõe de uma variedade de dados, com os quais pode experimentar livremente com análise de conteúdo, reconhecimento de padrões, comparações ou correlações. A partir dessa investigação, é possível identificar oportunidades de relacionamentos a serem explorados. As perguntas aparecem ao se olhar para os dados.

O surgimento do *Big Data* como conceito

A expressão *Big Data* reflete esse potencial e é mais recente, tendo surgido na literatura técnica por volta de 2005. Foi cunhada originalmente para descrever ambientes em que o volume de dados havia se tornado tão grande que não caberia em um computador, qualquer que fosse sua capacidade, e teria que ser trabalhado em muitos processadores em paralelo.

No entanto, rapidamente surgiram dois aspectos talvez mais relevantes dessa nova realidade. Um era a variedade de formas em que esses dados eram armazenados, não se adequando aos sistemas de gerenciamento de bases de dados existentes, que demandavam uma estruturação rígida e constante do conteúdo armazenado. O outro aspecto era o de potencial de uso desses dados: havia tanta informação, crescendo tão rapidamente, com tal variedade, que um novo desafio surgira, o de dar sentido e utilidade a essa informação toda. A nova abordagem era a de experimentar com dados, descobrir relações, "deixar os dados falarem por si" (MAYER-SCHÖNBERGER e CUKIER, 2013: 6-7).

Big Data, portanto, diz respeito ao que é possível se fazer com um volume de dados enorme, e que não estaria disponível ao se trabalhar com pequenos volumes de dados, abrindo novos horizontes:

> *"A maior parte das nossas instituições foram estabelecidas sob o pressuposto de que as decisões humanas baseiam-se em informação que é limitada, exata e causal por natureza. Mas tal situação muda quando o volume de dados é enorme, pode ser rapidamente processado e tolera a inexatidão. E devido ao vasto volume de dados, decisões serão tomadas não por humanos, mas por máquinas"* (MAYER-SCHÖNBERGER e CUKIER, 2013: 16) .

Uma curiosidade histórica sobre o termo *Big Data* é a de este ter surgido no campo da astronomia, não no das ciências sociais ou da administração. Na virada do século, foram construídos telescópios terrestres de grande capacidade, que captavam imensa quantidade de informações a partir das imagens do céu. Em 2000, o projeto Sloan Digital Sky Survey (SDSS), conduzido pelo observatório de Apache Point, no Novo México, EUA, acumulou em algumas semanas mais informações sobre o universo do que tudo o que havia sido registrado até então em toda a história da astronomia, algo como 1 TB de dados. Um de seus sucessores, o Large Synoptic Survey Telescope ou Observatório Vera Rubin, em El Peñon, no Chile, que deverá entrar em testes em 2022, irá colher essa mesma quantidade de dados na forma de imagens digitais em algumas horas de operação. Em um ano, esse volume de dados superaria 1 petabyte, demandando técnicas de mineração de dados para ser tratado, de modo a se chegar à catalogação exaustiva de todos os objetos celestes registrados[2].

Elementos básicos

Big Data, em suma, é o conjunto de técnicas que atendem a três objetivos:

- possibilitar o tratamento de volumes muito grandes de dados, grandes ao ponto de tornar inviável seu processamento em um único computador, por mais potente que este seja;
- como os dados são continuamente criados, inviabilizando em muitos casos seu armazenamento por questões de custo e de capacidade, deve ser possível tratá-los de imediato e descartá-los;
- dada a variedade de informações e de formatos, a abordagem é distinta da usada nos sistemas convencionais: propõe-se uma pergunta de pesquisa e testam-se exaustivamente as possibilidades de abordá-la a partir dos dados disponíveis ou, alternativamente, experimenta-se com os dados para identificar oportunidades de uso, ou seja, que tipo de pergunta poderia ser respondida.

Assim, dado o potencial inerente a um grande volume de dados, um conjunto de ferramentas de *Big Data* deve possibilitar uma abordagem investigativa, ou de análise preditiva, buscando antecipar a ocorrência de fatos ou de demandas. Esse é, talvez, o maior potencial de ganhos financeiros ou de posicionamento comercial para o setor privado.

Usualmente, um conjunto de ferramentas de *Big Data* deve atender às seguintes camadas de procedimentos:

- Coleta e acumulação de informações – envolve os aspectos de coleta de dados em grandes volumes e seu armazenamento, temporário ou permanente. Em decorrência do rápido crescimento dos repositórios,

[2] Comparativamente, o catálogo do SDSS contém cerca de 88 milhões de objetos celestes registrados.

aspectos de escalabilidade, ou seja, de acréscimo de recursos na medida da necessidade, são essenciais. Sistemas de gerenciamento de dados conhecidos como noSQL permitem estruturar esses dados em múltiplos formatos e armazená-los em vários locais, priorizando disponibilidade e escalabilidade[3].

- Multiprocessamento – é a infraestrutura da plataforma de processamento computacional e dos programas que viabilizam o tratamento em paralelo das informações desejadas; a solução de armazenamento, recuperação e tratamento de dados mais usada é o Hadoop, embora seja uma tecnologia "antiga"[4].
- Seleção e organização de dados – o uso de *Big Data* envolve a acumulação de informações diversificadas em grandes "lagos de dados", nos quais os dados são rastreados e recuperados para tratamento. O acesso a esses repositórios, no entanto, é lento para os padrões de processamento em tempo real que atualmente se impõem. Por esse motivo, dados devem ser buscados, padronizados, classificados, mantidos em memória e então trabalhados.
- Codificação de aplicações – aplicações de *Big Data* abordam problemas ou perguntas específicas que se desejam responder. Deve-se desenvolver um algoritmo, ou seja, uma "receita de bolo", que descreva que dados serão considerados, como serão selecionados e ajustados, como serão comparados e que critérios estatísticos serão aplicados para escolher relações válidas entre estes. Em alguns casos, a extensão e complexidade dos dados pode tornar excessivamente penoso ou demorado o processo de identificação de dados e de teste de alternativas. Eventualmente, técnicas de reconhecimento de padrões e de aprendizado computacional (*machine learning*) são incluídas na aplicação, para que o computador possa selecionar modelos ou tomar decisões a partir do comportamento dos dados. Linguagens específicas para elaborar *scripts* das operações a serem feitas, como Apache Pig ou Hive,

[3] Usualmente, para que essa flexibilidade seja alcançada, há um sacrifício de aspectos de segurança e consistência. Tais aspectos são centrais em modelos tradicionais adotados em sistemas de gerenciamento de bancos de dados (SGBD), como o modelo relacional do SQL, dificultando enormemente a distribuição dos dados em múltiplos locais. A categoria noSQL busca destacar essa distinção. Há dezenas de sistemas disponíveis, proprietários ou livres, podendo-se mencionar, entre os mais populares, o Apache Cassandra, o IBM Domino e o Couchbase.

[4] Hadoop foi desenvolvido pelo Yahoo! por volta de 2006, sobre o conceito de MapReduce desenvolvido pelo Google em 2003. Há soluções proprietárias ou livres mais recentes, como Spark, Storm ou BigQuery, com distintos recursos de tratamento e visualização de dados.

permitem programar algumas dessas aplicações de modo mais simples.

- Visualização de modelos, vistas e cubos – a supervisão do processo de escolha das correlações entre dados é feita por profissionais especializados na aplicação, que combinam a compreensão dos resultados estatísticos com sua própria intuição acerca da relevância dessas relações. Por isso, ferramentas de visualização gráfica dos dados, da forma como estão organizados e das relações identificadas, com recursos de detalhamento ou agregação das informações e de seleção de subconjuntos destas, são de grande eficácia.
- Aplicações de *business intelligence* (BI) e análise preditiva – a partir da relevância dos comportamentos e das correlações identificadas, é possível estabelecer regularidades nos dados, consolidar visões de negócios,. fazer antecipações de eventos ou de tendências que podem ser exploradas para a finalidade desejada. Pacotes estatísticos como R, SAS ou SPSS podem ser integrados nesses ambientes para prover ferramentas apropriadas a várias dessas aplicações.

Para cada uma dessas camadas de trabalho há uma variedade de ferramentas, além das que foram mencionadas acima. Em vários casos, estas são integradas em pacotes comerciais, como QlikTech, Tableau ou Spotfire. Em outros casos, empresas oferecem diretamente os serviços de análise de dados, a partir do seu conjunto de ferramentas, fazendo o trabalho computacional e deixando ao usuário apenas o esforço de supervisão e análise gráfica.

O mercado de *Big Data* vem se expandindo rapidamente em todas essas frentes. Apesar de ser um mercado ainda em desenvolvimento, já dispõe de soluções técnicas experimentadas e de empresas de consultoria e suporte consolidadas internacionalmente e com intensa atuação no Brasil. Segundo dados da Frost & Sullivan, o mercado brasileiro de *Big Data* e *analytics* movimentou em 2016 cerca de US$ 1,1 bilhão (COMPUTERWORLD, 2017).

3. Como prepara-se para trabalhar com *Big Data* e exemplos do que obter da técnica

Pesquisa do McKinsey Global Institute (MGI), em parceria com o McKinsey's Business Technology Office, realizada em 2011, apontava sete elementos-chave relacionados com a cultura empresarial que deveriam ser reconhecidos para explorar com eficácia o potencial oferecido pelo *Big Data* (MANYIKA et al, 2011):

- Preparar-se para explorar volumes de dados realmente grandes: cada empresa de mais de 100 empregados já teria, naquele ano, um acúmulo de 200 terabytes de dados armazenados.
- Ser capaz de tratar esses dados em cinco linhas distintas: elevar a transparência e usabilidade dos dados existentes,

estimular a variedade de dados coletados, aumentar a segmentação dos dados e a precisão das relações e tendências identificadas, investir em análise preditiva (*data analytics*) e incorporar a prática de análise de dados a todos os processos para buscar chaves para a inovação de produtos e serviços.

- Reconhecer que o uso de *Big Data* pode tornar-se essencial para a competição com concorrentes e a captura de valor. Saber explorar dados e obter intuições profundas e consistentes, em tempo real, valorizando o caráter estratégico da abordagem.
- Identificar e quantificar benefícios decorrentes do uso das ferramentas, que podem chegar a aumentos de margens operacionais das empresas de até 60%. Esses benefícios se estenderiam ao consumidor, na forma de ganhos de utilidade e satisfação com suas decisões de consumo. A comparação com os altos custos da técnica é indispensável.
- Reconhecer a possibilidade de um benefício assimétrico: haverá setores e atividades que terão grandes benefícios do uso do *Big Data*, mas outros terão menos ganhos. Saber qual a situação particular de cada caso é importante para a decisão de investimento. Os setores financeiro, de tecnologia da informação e de atividades governamentais eram apontados entre os que mais se beneficiariam.
- Administrar a dependência de talentos escassos. *Big Data* não envolve apenas profissionais e gerentes da área da computação, mas também pessoas com domínio da atividade fim e habilidade analítica de ponta. Essas pessoas têm que ser identificadas e preparadas.
- Preparar-se para uma revisão de políticas e regulações em vários setores: privacidade, segurança, propriedade intelectual e responsabilidade civil. Acesso a dados é um aspecto crítico para viabilizar essa ferramenta e isso tem dois enfoques distintos, pois externamente à empresa ou entidade os elementos de legislação e mercado para tal devem ser consolidados e, internamente, normas e práticas devem ser alinhadas.

Iniciativas organizacionais

Tais elementos-chave põem em destaque um aspecto que vai além do domínio dos conceitos e ferramentas de *Big Data*. O interessado em desenvolver-se nessa direção terá que se organizar para isso.

De um ponto de vista estratégico, há três aspectos importantes: definir objetivos e perguntas de pesquisa relevantes para trabalhar o acervo de dados (diretriz estratégica), preservar a consistência, integridade e adequação dos dados (ativo estratégico) e, enfim, preparar-se para enfrentar os custos envolvidos (investimento estratégico).

A gestão de ativos estratégicos em computação tende a estar focada na infraestrutura montada pela empresa e na contratação de serviços a terceiros, ou seja, uma visão que privilegia os meios que devem ficar à disposição dos vários usuários para que suas necessidades sejam atendidas. No entanto, *Big Data* requer uma gestão do acervo de dados usado pela empresa ou entidade. Isto inclui não apenas os dados próprios, mas também aqueles coletados ou contratados junto a terceiros.

A gestão de dados é importante por vários motivos. Primeiro, porque o estoque de dados que poderá trazer retornos à empresa ou entidade é muito grande, é em parte não documentado e há um risco recorrente de perdas de dados. Desse modo, procedimentos de acompanhamento desse "lago de dados" devem ser estabelecidos.

Outro motivo é o de que o acesso e o uso desses dados deve ser realizado com a rapidez inerente às demandas de pesquisa ou descoberta que estão por trás dos projetos de *Big Data*. Assim, dados que sejam estáticos e cuja utilização seja previsível devem ser formatados para compor um repositório de fácil acesso (*data warehouse*).

Um terceiro motivo para uma gestão dos dados é o de que tratar com *Big Data* envolve pessoas especializadas. O trabalho requer uma combinação de capacidades bastante diversificada, de modo que se administram talentos escassos. Um praticante qualificado deve combinar habilidades de informática (compreender a arquitetura e as ferramentas de *Big Data*, programar com competência), de estatística (análise estatística, análise visual, identificação de padrões, compreensão do processo de *machine learning*) e de domínio da aplicação (conhecimento de negócio, compreensão das demandas e da tecnologia envolvidas, tomada de decisões). Será preciso identificar, preparar e preservar esses profissionais, bem como montar e administrar equipes específicas para os projetos de *Big Data*, tipicamente multidisciplinares.

Nas empresas, isto resultou no surgimento de toda uma classe de especializações e denominações profissionais antes inexistente: cientistas de dados, analistas de negócios, especialistas em segurança de dados, engenheiros de *warehousing*, especialistas em mineração e *data analytics*.

O problema dos custos dos dados, que detalharemos mais andiante, é, enfim, mais um elemento para que a gestão de dados deva ser eficaz. Contratar dados é um processo caro. Expandir capacidade operacional para preservá-los, também. Uma gestão eficaz pode determinar as condições em que cada base de dados deve ser preservada, para ajustar esses custos às perspectivas de uso desses dados e ao retorno esperado.

Algumas aplicações e histórias de sucesso

(1) Serviços de catalogação na internet

Os serviços de catalogação na internet passaram por uma revolução com o uso de grandes volumes de dados. A criação do Google, em especial, representou uma profunda transformação nesses serviços (LINS, 2013: 48-49: ISAACSON, 2014: 461).

Catálogos de sites não eram novidade. Em 1994 já existia o Yahoo!, um guia temático da web construído a mão. Também começaram a surgir os programas de navegação automática (*web crawlers*), que passeavam pelas páginas seguindo os hyperlinks existentes e desse modo construíam catálogos automáticos. Serviços como Alta Vista, Lycos ou Infoseek usavam essa estratégia. O próprio Yahoo! passou a combinar os resultados de seu próprio sistema *crawler* com a organização temática que o caracterizava.

A diferença que o Google trouxe foi a combinação de uma interface muito simples com um modelo inovador de extração de dados da web, baseado em mineração de dados. A ideia era oferecer ao usuário do serviço uma lista de páginas que atendessem a um termo de pesquisa em uma ordem que refletisse o valor ou mérito de cada página. Páginas mais importantes viriam antes. O primeiro critério de relevância seria o de quantos links apontariam para a mesma, algo fácil de definir, mas dificílimo de quantificar.

O algoritmo do Google acumulava um volume de dados impressionante a respeito desses milhões de links para estimar, periodicamente, esse valor para cada site visitado. Os programadores do Google, Larry Page e Sergey Brin, e suas equipes, refinaram com o passar dos anos os critérios de avaliação das páginas e a capacidade de armazenamento do serviço, oferecendo resultados muito superiores aos de outros mecanismos de busca. Em 1998, pouco tempo após entrar em operação, a base de dados do Google já continha mais de 500 milhões de hyperlinks, dos três bilhões que, estima-se, existissem então na web.

Esse exemplo, mais uma vez destaca a diferença entre uma abordagem convencional, usada por outros serviços de busca, e a abordagem de mineração de dados, usada pelo Google.

Mais do que facilitar a vida do usuário, o Google consolidou uma nova forma de interagir com a internet. Não era mais necessário escolher um site e iniciar um passeio pela rede. Era possível consultar o Google mediante uma palavra chave ou uma expressão de busca e obter o resultado mais relevante naquele momento, para ir diretamente aonde se desejava. O Google revelou ser mais do que um catálogo da rede: era sua porta de entrada.

O alcance dos serviços do Google e sua capacidade de extrair novos resultados e criar novos serviços a partir de sua imensa coleção de dados não para de surpreender, como ilustra o exemplo já apresentado na seção 2 deste texto.

(2) Orientação de campanhas eleitorais

Esta talvez seja a linha de aplicações do *Big Data* que tenha recebido maior divulgação da imprensa nos últimos anos, pela variedade de casos de sucesso que vieram a público. A campanha em defesa do Brexit, no Reino Unido, em 2016, serviu como caso de referência para a adoção de *Big Data* na política. Sua eficácia baseou-se na tese de que é possível, a partir do comportamento de uma pessoa nas redes sociais, estabelecer algumas

medidas acerca de seus traços psicológicos e sua personalidade. Trata-se de uma técnica de trabalho denominada psicometria (LINS, 2019: 289-292)..

São colhidos grandes volumes de informação a respeito das ações de cada pessoa na internet: a que redes sociais se conecta, quais as características de suas "curtidas", suas buscas, a frequência de suas mensagens e assim por diante. São informações de comportamento, não havendo quebra do sigilo de postagens ou uso de informações pessoais.

Ainda assim, a partir de informações tão superficiais, é possível, graças ao grande volume, chegar a medidas bastante precisas de traços da personalidade, como o quanto essa pessoa é extrovertida, perfeccionista, afável, emocionalmente equilibrada e aberta a novas ideias, cinco grandes componentes do seu comportamento, dimensões conhecidas como OCEAN (dos nomes em inglês, *openness*, *conscientiousness*, *extraversion*, *agreeableness* e *neuroticism*). Em uma abordagem convencional, essas dimensões seriam estimadas a partir de questionários bastante extensos respondidos pela pessoa. O "pulo do gato" foi dado por pesquisas na Universidade de Cambridge, no Reino Unido, que conseguiram correlacionar essas dimensões com dados do comportamento das pessoas nas redes sociais. Além disso, é possível derivar dessas informações alguns atributos pessoais, como padrões de consumo, grupo socioeconômico a que pertence a pessoa, preferências políticas, afiliação religiosa ou opção sexual (KOSINSKI et al, 2013).

Desse modo, mapeia-se onde essas pessoas estão, quais suas preferências, quais os seus temores, que palavras-chave as fazem tomar decisões. O efeito das informações psicométricas no planejamento de campanhas eleitorais fez-se rapidamente sentir. Esses dados podem ser usados tanto para orientar uma abordagem ao eleitor quanto para direcionar mensagens que o sensibilizam.

Os candidatos que adotaram a técnica mudaram sua forma de atuar na campanha. Em lugar de buscar o uso intenso da mídia tradicional como canal de informação do público e de eventos e convenções para mobilização partidária, ajustamdo o discurso a uma posição capaz de angariar o maior número possível de eleitores, passaram a identificar grupos de eleitores na internet com perfis e interesses parecidos entre si, direcionando-lhes mensagens específicas. O discurso do candidato reforça uma variedade de pontos muito focados, muitas vezes sem conexão aparente, que atendem ao sentimento dessas parcelas de público específicas. A técnica mostrou-se mais eficaz no acesso ao eleitor e no seu convencimento.

Historicamente, as primeiras campanhas de expressão em que essas técnicas foram adotadas, ainda que de forma incipiente, foram as de Barak Obama à presidência e à reeleição, nos EUA. Posteriormente, um crescente número de campanhas vitoriosas apoiou-se nessa abordagem, entre as quais a já citada do Brexit, as candidaturas de Donald Trump e, no Brasil, de João Dória e Jair Bolsonaro. Também fizeram uso desse ferramental os empreendimentos eleitorais de Mauricio Macri, na Argentina, e de vários candidatos nas eleições gerais de países europeus nos últimos anos, entre os quais Emmanuel Macron e Marine Le Pen na França.

(3) Investigação criminal e inteligência

Aplicações de *Big Data* e análise preditiva começaram a ser adotadas na investigação criminal e em atividades de inteligência a partir do episódio de 11 de setembro de 2001. Logo após a queda das torres gêmeas, houve evidências de que referências a respeito da possibilidade do atentado existiam, de modo esparso, em informes recebidos pelas agências norte-americanas de inteligência. No debate público que se seguiu, surgiu a expressão "o desafio do volume" (*volume challenge*). A comunidade de inteligência norte-americana vinha se debatendo com o crescente volume de dados havia anos, sem encontrar um caminho para lidar adequadamente com isso e convivendo com a falta de coordenação entre quinze agências de inteligência distintas (ADAMS et al, 2011: 15; McCUE, 2015: 37).

Os resultados da comissão indicada para examinar as falhas associadas ao ataque levaram o congresso dos EUA e o presidente George W. Bush a sancionar o Patriot Act um mês após os ataques. Entre as medidas de segurança doméstica, vigilância e investigação previstas, foi admitido o monitoramento de comunicações de cidadãos norte-americanos, inclusive na internet, e a identificação de seus padrões de comportamento. Isto resultou em projetos da NSA de coleta maciça de dados sobre comunicações e na tentativa de identificar correlações com eventos suspeitos.

A extensão e a capacidade de processamento dos projetos assim conduzidos são, em grande medida, desconhecidos. No entanto, algumas evidências indiretas, como as revelações do administrador de sistemas Edward Snowden, que divulgou, em 2013, um significativo volume de documentos relacionados com as práticas de monitoramento da rede mundial, sugerem que o volume de coleta de dados pode ser espantosamente alto, alcançando bilhões de transações por hora. Em entrevista publicada pela revista Wired, ressalta-se:

> "*[Snowden] começaria a constatar o enorme alcance da capacidade de vigilância da NSA, uma habilidade de mapear o movimento de qualquer morador de uma cidade monitorando seu endereço MAC, um identificador único emitido por cada celular, computador ou outro equipamento eletrônico. (...) Ele pode confirmar, segundo afirma, que grandes volumes de comunicações dos EUA vinham sendo interceptados e armazenados sem qualquer mandado, sem um motivo de suspeita criminal, motivação provável ou identificação individual*" (BAMFORD, 2014: 92, 93)

Infelizmente, como destacam Adams et al (2015: 18-20), os esforços de mineração de dados conduzidos dessa forma não tiveram eficácia comprovada quando submetidos ao escrutínio das autoridades norte-americanas, provavelmente porque atividades terroristas são eventos raros e não houve como detectar padrões com significância estatística nesses casos.

Por outro lado, a mineração de dados revelou-se uma técnica interessante para a investigação de crimes comuns e vem sendo usada com crescente regularidade. Várias correlações inesperadas foram identificadas nas bases de dados sobre crimes e agressões das polícias estaduais norte-americanas. Por exemplo, determinou-se, na investigação de casos de violência sexual, que ocorrências anteriores de furtos residenciais tinham maior correlação com o cometimento de um estupro do que ocorrências anteriores de agressão sexual. O uso desse tipo de correlação estendeu o espectro de investigação sobre suspeitos de casos específicos em que amostras de DNA haviam sido colhidas e poderiam ser usadas como prova.

(4) Acompanhamento e avaliação de políticas públicas de saúde

Trata-se de outra área que se expande rapidamente, em especial após a pandemia COVID-19. Os governos tendem a acumular grandes volumes de informação pessoal dos cidadãos, que podem ser descaracterizadas (anonimizadas), eliminando a identificação pessoal, e usadas para fins de pesquisa e de concepção ou acompanhamento de políticas públicas.

Não é uma abordagem nova no setor público. No Brasil, por exemplo, os censos da população e as pesquisas nacionais por amostragem de domicílios, as PNAD, já incorporam perguntas diversas sobre conhecimento de políticas públicas e apropriação de seus benefícios. A mesma metodologia é aplicada nas pesquisas voltadas às empresas. Esta é, no entanto, uma abordagem convencional, com o retardo inerente à sua condução, pois demanda o acesso ao agente ou cidadão, o preenchimento do relatório, a transcrição e a catalogação das informações.

Uma abordagem alternativa pode ser buscada ao se tratar de registros que são mantidos de modo contínuo no relacionamento do cidadão com o Estado. Murdoch e Detsky (2013: 1351) apontam, por exemplo, que a adoção de prontuários eletrônicos nos EUA (*electronic health records* – EHR) deu origem a um vasto repositório de informações sobre a saúde da população. Em parte, esses prontuários são preenchidos pelos médicos com relatos descritivos da saúde dos pacientes e o objetivo original da iniciativa era o de garantir que, ao serem atendidos em uma unidade, já se dispusesse de imediato do histórico clínico destes.

O volume e o detalhe dessas informações são enormes. E vários usos potenciais estão sendo antecipados, de modo que esse estoque de dados, até recentemente considerado um custo colateral ou um refugo do sistema, passou a ser um ativo estratégico.

Uma aplicação já em curso é a identificação de antecedentes que possam ser vinculados a algum episódio clínico, caracterizando comorbidades, ou, em uma abordagem preventiva, a identificação de pacientes que possam vir a sofrê-lo. Andreu-Perez et al (2015: 1195) apontavam que estudos epidemiológicos já começavam a explorar esse estoque de dados há alguns anos para identificar situações desse tipo.

O uso de ferramentas de *Big Data* já se encontra bastante disseminado nas ciências médicas e na bioquímica, devido à complexidade e ao volume de dados que é preciso tratar para desenvolver certos projetos em bioengenharia e farmacologia. Por exemplo, um sequenciamento de genoma, dependendo do grau de detalhe desejado, envolve um volume de dados de até 200 GB (ANDREU-PEREZ et al, 2015: 1197). Esse tipo de informação pode ser útil para mapear a suscetibilidade a doenças, examinar interações farmacológicas ou a eficácia de tratamentos, mas requer o acompanhamento de grande número de pacientes, caracterizando portanto um problema típico de *Big Data.*

4. Custos envolvidos no *Big Data*

Como foi anteriormente explicado, *Big Data* refere-se não apenas ao tratamento de grandes volumes de dados, mas também à integração ou correlação entre dados de natureza e formatos distintos, com recursos de análise para produzir um tipo de conhecimento prospectivo. Deseja-se responder não apenas a perguntas do tipo "o que está ocorrendo" ou "qual o vínculo de causalidade entre isto e aquilo", mas também questões como "e se eu oferecer tal opção..." ou "que tal se...". Buscam-se, em suma, opções cujo resultado é incerto e cuja relação de causalidade é desconhecida, mas que trazem potencial de ganhos expressivos.

Parte dessas possibilidades estão agregadas sob o rótulo da tecnologia social. A interação entre pessoas nas redes sociais e sua receptividade a informações de empresas e governo são elementos importantes para a oferta de produtos, serviços e políticas públicas mais eficazes. Além disso, como a interação das pessoas nas redes é intensa, revela muito sobre seu comportamento.

O *Big Data* incorpora recursos para o acompanhamento e a análise dessas interações. Como comentado na seção anterior, o modo como o internauta acessa a rede, o tipo de decisão que toma, desde uma aprovação de uma mensagem ou a postagem de um texto até o compartilhamento de imagens e a busca de contatos, está ligado a valores fundamentais de sua personalidade. Entender como as relações entre seus atos e seus valores se configuram pode ajudar a oferecer produtos, serviços ou informações que sejam mais bem recebidos e eficazes.

Muitos dos recursos da tecnologia social e do seu monitoramento podem ser implementados para comunicação interna e ganho de desempenho em comunidades fechadas, tais como grupos de interesse, funcionários de empresas ou parceiros em negócios. Ganhos de produtividade de até 25% têm sido observados em alguns desses casos (CHUI et al, 2012).

Tratar dados dessa natureza e com esse enfoque requer abordagens e ferramentas específicas. Vamos relembrar que há cinco problemas no processamento desses dados que não podem ser resolvidos com uma abordagem convencional e envolvem custos importantes. Graças ao avanço da tecnologia da computação, porém, novas ferramentas foram desenvolvidas e, em vários casos, são recursos de software aberto, facilmente disponíveis. No entanto, envolvem custos expressivos.

Esses cinco problemas próprios do *Big Data* já foram comentados e a seguir os repisamos:

- Os dados não estão prontamente disponíveis. É preciso descobri-los e, em muitos casos, adquiri-los.
- Os dados envolvem volumes muito grandes e não há como processá-los em um único computador, por mais poderoso que seja. Seu tratamento demanda o uso de múltiplos equipamentos e esse processamento tem que ser escalável, ou seja, a bancada de processadores deve aumentar ou diminuir na medida do tamanho do conjunto de dados a ser tratado.

- Os dados têm formatos distintos e heterogêneos. É preciso trazê-los a um contexto que permita seu tratamento.
- Os dados em muitos casos não são armazenados, pois são criados continuamente e sua acumulação envolveria uma capacidade de armazenamento caríssima. São, então, tratados em tempo real e descartados. O software precisa determinar dinamicamente a melhor forma de correlacioná-los e processá-los diretamente na memória dos computadores envolvidos.
- Os dados devem ser compreendidos e trabalhados por pessoas não especializadas em computação, mas qualificadas na atividade fim que fará uso desses resultados. Desse modo, é necessário um mecanismo para apresentar graficamente os resultados e permitir a escolha de graus variados de segmentação.

Em parte, esses desafios estão embutidos no que se conhece como os "três V" do *Big Data*: volume, velocidade e variedade (McAFEE e BRYNJOLFSSON, 2012: 63).

Custos do processamento escalável

O enorme volume de dados envolvido impede que estes sejam tratados em um único processador. É preciso que se ofereça uma arquitetura e um mecanismo de processamento escalável, com o uso de muitos processadores. Como já se comentou anteriormente, as soluções mais utilizadas hoje baseiam-se em uma arquitetura de processamento desenvolvida pelo Google, chamada MapReduce, e um ambiente de armazenamento do Yahoo!, o Hadoop.

Custos de armazenamento e processamento não são importantes quando examinados por unidade de informação armazenada. Na realidade, os custos de equipamentos diminuíram significativamente ao longo das últimas décadas e hoje é possível dispor de alternativas que ocupam pouco espaço, consomem pouca energia e permitem, em seu conjunto, tratar um volume de dados significativo. O custo de software pode ser também administrado com a adoção de programas de livre distribuição (*open software*), e há algumas boas alternativas disponíveis no mercado.

O problema, no entanto, é o de que os volumes de dados armazenados são de fato muito elevados. Ademais, a busca de correlações entre estes pode consumir uma capacidade de processamento muito grande. Há um problema não apenas de "*big data*", mas também de "*big analytics*". Enquanto uma empresa de médio porte mantém uma base de dados de algumas dezenas de terabytes de tamanho, empresas que se apoiam fortemente em dados mudam rapidamente de escala. McAfee e Brynjolfsson (2012) mencionam, por exemplo, que o Walmart coleta mais de 2,5 petabytes[5] de dados por hora

[5] 1 terabyte (TB) equivale a 1.024 megabytes (MB). 1 petabyte (PB) equivale a 1.024 terabytes.

sobre as transações com clientes. Desse modo, custos de armazenamento e processamento disparam.

Ferramentas estatísticas e *data analytics*

O volume de dados envolvido em *Big Data* impede em muitos casos seu armazenamento definitivo. Ferramentas como Hadoop viabilizam a retenção pelo tempo necessário para sua avaliação e tratamento. A velocidade de renovação dos dados, porém, faz com que esse tratamento fique limitado em muitos casos a um nível rudimentar de análise.

Há uma distinção importante entre o tratamento analítico de dados de formato predeterminado e as possibilidades abertas pelo *Big Data*. Quando o formato dos dados é uniforme e previamente conhecido, como é o caso de bases de dados convencionais, o analista sabe com exatidão que tipo de informação estará disponível e pode antecipar e projetar as possíveis correlações que poderão ser estabelecidas, em geral a partir de um marco teórico preexistente. Para cada uma dessas correlações define-se que hipótese se deseja testar. O objetivo da análise é descartar ou aceitar essa hipótese.

Já no caso do *Big Data*, não há um conjunto de dados previamente estruturado a utilizar. O analista está permanentemente envolvido com o esforço de encontrar novos dados e intuir algum uso interessante que se possa fazer destes. As correlações não são concebidas somente a partir de uma teoria preexistente mas, em muitos casos, surgem de uma experimentação por tentativa e erro.

A sequência de trabalho, então, é oposta nos dois casos. Com dados estruturados, há um modelo prévio que determina a estrutura da base de dados, a coleta das informações, sua seleção e consolidação. Com *Big Data*, a escolha e a observação do comportamento dos dados é o que leva o analista a levantar hipóteses e testá-las, construindo um modelo ou um conjunto de conjecturas a partir de correlações bem sucedidas. Portanto, o uso de *Big Data* é intensivo em trabalho e requer grande esforço de atenção do praticante.

É possível fazer alguma iniciativa em *Big Data* a um custo de análise moderado. Pode-se fazer processamento paralelo com computadores padrão. Alguns procedimentos simples de apresentação e segmentação de dados podem ser operados com macros sobre planilhas ou sistemas de gerenciamento de banco de dados de baixo custo. Certos softwares estatísticos, como o R, são de livre distribuição e dispõem de boas bibliotecas de ferramentas analíticas. No entanto, o aumento no tamanho do problema ou da variedade de dados a tratar rapidamente eleva os custos.

Agregue-se que, mesmo em casos simples, o esforço de trabalho é elevado e o custo de dedicação de equipes de analistas de dados é alto. Não há como reduzir significativamente esse elemento de custo.

Em suma

Na medida em que se avança no uso e na sofisticação da aplicação, os custos disparam por três motivos. O primeiro é o de que dados são caros. Com raras exceções, o acesso a repositórios ou a bases de dados não é gratuito. Deve-se pagar por estes, e pagar caro. E deve-se armazenar dados em volumes elevados. Quanto mais detalhados e

personalizados forem esses dados, maior a conta. O segundo é o do que determinar correlações entre dados, em especial quando não se assume um modelo de causalidade, é um exercício exaustivo de testar todas as possibilidades disponíveis. Isto envolve o cruzamento de centenas ou milhares de alternativas para se selecionar aquelas que tragam algum resultado estatisticamente significativo. O processamento desses dados, portanto, exige uma capacidade computacional maciça. O terceiro motivo, enfim, é o de a qualidade das correlações obtidas depende do nível de desagregação, ou de detalhe, dos dados usados. Isto demanda um exercício de identificar possíveis oportunidades e trabalhá-las em detalhe, consumindo o tempo de especialistas e elevando os custos de pessoal.

Portanto, o custo do *Big Data* é inevitavelmente um ponto fraco na adoção dessa tecnologia. Ela fará sentido quando o retorno esperado, seja em ganhos pecuniários, seja em ganhos sociais, venha a justificar esse pesado investimento. Por enquanto, *Big Data* ainda não é para todos.

5. *Big Data* e gestão do empreendimento

As aplicações bem sucedidas de *Big Data* que chegam ao público ilustram sua adequação a estudos de mercado e a iniciativas de caráter coletivo ou de políticas públicas. As oportunidades de uso no ambiente da empresa são igualmente interessantes.

Trabalhar com dados e fatos tem sido uma diretriz na estruturação e administração da produção e de negócios, seja no ambiente fabril, seja em atividades de campo e em serviços. O mapeamento dos processos, a identificação de indicadores-chave para seu acompanhamento e a tomada de decisões a partir da identificação de situações em que o processo está fora de controle fazem parte do rol de atividades mais tradicional da engenharia da qualidade. O que há de novo, então?

O volume de dados coletados no ambiente de produção multiplicou-se em decorrência de algumas tendências: a automação industrial, o monitoramento de equipamentos e instalações a partir de dados de sensores e a montagem de redes de supervisão com sensores e atuadores. Esses dados, além de servirem aos objetivos para os quais foram originalmente colhidos, podem ser aplicados a outras finalidades.

Dados de sensores instalados em equipamentos, por exemplo, servem para o controle de servomecanismos e para o acompanhamento de desgaste, de ajustes e de níveis de serviços, indicando o momento de realizar manutenções preventivas. A coleta ampla desses indicadores, por outro lado, propicia a oportunidade de agregar essas informações a sistemas de controle da produção, de modo a administrar a carga de uso dos equipamentos e controlar os efeitos sobre os prazos de manutenção, otimizando a operação conjunta e minimizando custos e paradas.

Na gestão do empreendimento, as oportunidades para o uso de grandes massas de dados tornam-se a cada dia mais frequentes. Uma anedota citada em diversos artigos sobre o tema ilustra o potencial do *Big Data* em atividades corporativas, relacionada com as lojas de moda Target (HILL, 2012).

A grife havia desenvolvido um estudo de *Big Data* que correlacionou alguns de seus dados, como o bairro em que suas clientes moravam, sua idade, e o consumo

de duas dúzias de produtos de seu estoque, à probabilidade de estarem grávidas e serem suscetíveis a ofertas de roupas infantis ou fraldas descartáveis. Em certos casos, era possível estimar inclusive o trimestre de gravidez da cliente. A partir dessas inferências, a loja passara a enviar cupons às clientes potencialmente grávidas, como parte de uma política de fidelização. A anedota refere-se a um senhor da cidade de Minneapolis que reclamou ao gerente de uma loja que sua filha adolescente vinha recebendo essas correspondências: "Vocês querem encorajá-la a engravidar?". O gerente, alarmado, retirou o nome da jovem da mala direta para, alguns dias depois, receber um acanhado pedido de desculpas por telefone. A jovem estava de fato grávida e não havia dito nada aos pais.

Note-se, mais uma vez, que não havia um estudo prévio que pudesse estabelecer uma causalidade entre o conjunto de dados utilizado e as conclusões a que a loja chegava a respeito de suas clientes. A aplicação utilizada decorria apenas de testes com grandes volumes de dados e, no caso em particular, é improvável que um modelo comportamental pudesse ser estabelecido para explicar as correlações obtidas. Mas estas funcionavam.

A anedota, além de ilustrar o alcance das análises sobre grandes volumes de dados, traz à tona um outro aspecto: os ganhos potenciais dos próprios clientes, na forma de ganhos de utilidade com sugestões de pesquisa em portais, oferta de descontos em produtos ou serviços, recomendações de contatos e assim por diante.

No âmbito público, os ganhos podem ser igualmente relevantes. Ao ajustar políticas públicas e examinar seu retorno efetivo, órgãos de governo podem dar mais eficácia a iniciativas importantes e de elevado custo, a exemplo de campanhas de prevenção de doenças, auxílio no aprendizado de crianças ou melhoria da qualidade do policiamento público.

O problema dos dados pessoais

Diversas aplicações desenvolvidas por empresas tratam de dados descritivos de pessoas. Dados para contato, indicadores agregados ou desagregados de comportamento nas redes sociais, resultados de pesquisas de opinião, desempenho de interações com clientes, compra de cadastros são algumas das possíveis fontes de informações pessoais utilizadas com regularidade.

O uso de informações pessoais envolve um grau de responsabilidade, inclusive quando ocorre no contexto de grandes massas de dados sendo trabalhados de forma prospectiva. Em decorrência das preocupações que foram se acumulando ao longo dos anos acerca de possíveis constrangimentos às pessoas e invasões da privacidade, uma legislação internacional vem sendo consolidadada e esta envolve responsabilidades e custos importantes para a empresa.

Um marco importante para o tratamento de dados pessoais foi a adoção da Convenção de Estrasburgo de 1981, cujo art. 2º define dado de caráter pessoal como "qualquer informação relativa a uma pessoa singular identificada ou suscetível de identificação ('titular dos dados')". Dados de caráter pessoal, segundo o art. 5º da Convenção,

devem ser obtidos e tratados de forma leal e lícita, registrados com fim determinado e conservados por período que não exceda à finalidade do seu registro[6].

Na legislação brasileira, a Lei nº 13.709, de 14 de agosto de 2018, Lei Geral de Proteção de Dados Pessoais (LGPD), considera dado pessoal a informação relacionada a pessoa natural identificada ou identificável. A regulamentação do Marco Civil da Internet, Decreto nº 8.771, de 11 de maio de 2016, define como dado pessoal o dado relacionado à pessoa natural identificada ou identificável, inclusive números identificativos, dados locacionais ou identificadores eletrônicos, quando estes estiverem relacionados a uma pessoa.

Os desafios para o tratamento de dados pessoais com *Big Data* envolvem aspectos da tecnologia discutidos anteriormente. Lembre-se que *Big Data* pode servir para se correlacionar informações aparentemente desconexas e estabelecer perfis (o exemplo de psicometria aplicada às eleições é ilustrativo). Nesse caso, são acessados, por exemplo, dados comportamentais de observação pública, como os "likes" em rede social que o indivíduo submeteu, para destes derivar presunções a respeito de aspectos mais profundos da sua personalidade ou de suas preferências. Esses dados não foram registrados ou colhidos junto ao titular, mas construídos. As relações estabelecidas entre estes não são determinadas, mas probabilísticas. Não são do conhecimento do interessado, mas inferidos por um terceiro.

O enquadramento desses dados no âmbito dos dados pessoais é evidente. De fato, referem-se a pessoa determinada, afetam sua privacidade e não foram objeto de discussão pública. No entanto, como pode alguém deter o controle sobre uma informação a seu respeito que ele não forneceu, que, em certa medida, foi inventada, e da qual ele nem mesmo tem conhecimento?

Há uma ampla discussão sobre a aplicação ou não dessa problematização a dados anonimizados, ou seja, sobre os quais foi aplicado algum procedimento de retirada do vínculo com o indivíduo a que se referem, de agregação de dados, de vinculação a pseudônimo ou a informação indicativa de outra referência, como o terminal móvel em uso, ou de mescla ou reordenação de dados. Dados anônimos ou submetidos a um processo de descaracterização desse tipo estariam fora do alcance de um titular. O conceito é controverso no âmbito de aplicações de *Big Data*, pois o cruzamento entre grandes bases de dados não conexas pode propiciar mecanismos de identificação do indivíduo descaracterizado ou de seu enquadramento em um perfil que possibilite sua ulterior localização.

A LGPD é criteriosa quanto ao tratamento de dados pessoais, impondo um regramento à empresa que faz uso destes, limitando sua aplicação aos fins para os quais foram colhidos e impondo um ônus burocrático de registro junto a uma autoridade pública.

Pessoas notórias podem beneficiar-se de forma contratual de uma precificação de dados pessoais e informações a seu respeito. Da cobrança de cachês por

[6] O alcance do conceito de dado pessoal está acentuadamente relacionado a esse aspecto de uma definição mais restritiva, associando o dado a pessoa claramente identificada, ou mais expansionista, reconhecendo o dado como associado a pessoa de algum modo identificável (MACHADO et al, 2015: 17-18).

entrevistas e exposição à remuneração de publicidade em suas páginas, há uma variedade de modalidades de arrecadação a seu dispor.

Já as pessoas comuns não têm acesso a esses mercados. No entanto, "vendem" suas informações, às vezes sem dar atenção ao fato, ao aceitar regras de acesso a aplicativos e redes, ao contratar serviços ou ao fazer uso de benefícios públicos. Seus cadastros e as operações que realizam são integrados a bancos de dados com graus variados de sigilo, que eventualmente poderão ser cruzados em algum momento sem seu conhecimento. E, em muitos casos, foi dada inadvertidamente uma autorização ampla para o livre uso desses dados. Há, pois, espaço para reclamações e, apesar do amadurecimento da legislação, a empresa que usa esses dados permanece exposta a riscos.

O ajuste do tratamento de dados pessoais deve, portanto, calibrar esses aspectos. A empresa deve delimitar com clareza o alcance da coleta e uso de dados pessoais e os critérios éticos adotados, além de atender aos requisitos legais. Por outro lado, deve ser capaz de justificar e preservar sua aplicação nos casos em que a titularidade seja descaracterizada ou em que haja oportunidade de benefícios ao próprio usuário. Além disso, deve considerar o aspecto de que exista conhecimento disponível a respeito das pessoas baseado em informações pessoais de fluxo, para as quais não exista, necessariamente, armazenamento de dados específicos.

Atribuição de responsabilidades e *machine learning*

No tema da responsabilidade sobre a manutenção de dados, sua segurança, sua utilização e seu fornecimento no caso de investigações, um aspecto complicador adicional é dado pelo uso de algoritmos de *machine learning* em muitas aplicações, especialmente quando se trabalha com dados de fluxo, ou seja, aqueles que são usados de imediato e não são preservados.

O desafio dessa técnica é o de que em muitos casos o computador é programado com soluções que não oferecem condições de acompanhamento externo da lógica que é construída a partir da exposição aos dados. Um exemplo é o das redes neurais, que vão calibrando sua decisão na medida dos erros e acertos registrados, ou seja, há uma espécie de "adestramento" ou "calibração" incorporado ao software.

A descrição que acabamos de propor infelizmente é algo obscura, ilustrando um desafio de quem escreve sobre computação: a compreensão profunda de métodos e técnicas dependeria de um detalhamento relativamente pormenorizado. Redes neurais, porém, não são novidade. Seu uso encontra-se amplamente incorporado a uma variedade de aplicativos, tais como reconhecimento de escrita em digitalização de documentos, tratamento digital de imagens, manobra automática de veículos, buscas em bancos de dados, distribuição de correspondência, compensação de cheques, operação automática de corretagem em bolsas de valores e assim por diante.

Alguns desses casos podem envolver questões de responsabilidade civil, abrindo um problema importante, pois a decisão automática foi feita sob parâmetros que escapam ao completo controle seja do programador, seja do profissional da aplicação.

O aspecto se agrava no caso de sistemas especialistas. São programas que emulam a coleta e seleção de dados, a análise e a tomada de decisões que

pessoas especializadas fariam diante de um processo de trabalho. Usualmente dedicados a aplicações muito específicas, esses sistemas tomam o lugar de pessoas por serem capazes de tomada de decisões em grande velocidade sobre dados em fluxo, uma habilidade que o profissional humano não possui. Por outro lado, sistemas especialistas não são isentos de falhas ou imprecisões, devendo ser supervisionados com regularidade. Muitas aplicações desse tipo estão integradas a ambientes de manufatura ou a operações em campo, por exemplo na indústria de mineração. Em diversos casos, esses sistemas são adotados para indicar a necessidade e conveniência de realizar ações de inspeção, de calibração ou de manutenção preventiva, selecionando as situações sujeitas a um acompanhamento por pessoas.

Essa circunstância exige do profissional de controle da produção e do profissional da qualidade uma visão diferente a respeito da sua atividade, sobretudo quando envolve a supervisão desses sistemas. Ele não acompanha detalhes de cálculos, tomada de decisões ou definição de diretrizes. Sua formação deve ser maia refinada, para que ele possa construir um parecer e supervisionar todo o processo, identificando episódios dúbios ou que possam representar perdas para a empresa ou riscos para os empregados, para clientes ou para a comunidade, sem a necessidade de embrenhar-se em pormenores.

Em suma, o sistema especialista não é o responsável técnico e não assina uma ART. A responsabilidade do engenheiro e do gestor continuam a ser inalienáveis.

Desafios éticos

Há um mantra nos estudos de tecnologia, que nos diz que "tecnologia é neutra". Esta pode ser usada para o bem ou para o mal, dependendo de quem a aplica. No caso das soluções de computação, trata-se de uma meia verdade. As ferramentas de software em geral só fazem sentido diante de uma aplicação concebida, desenhada, programada e calibrada. Os princípios, valores, preconceitos e táticas de quem a usa estão, portanto, incorporados à solução. Ao adquirir um produto ou serviço que incorpora software estamos, em alguma medida, nos colocando no espaço ético de quem o desevolveu.

As narrativas de situações em que problemas éticos se configuram devido ao uso de *Big Data*, de ferramentas e serviços de *data analytics* e de sistemas especialistas, mostram um quadro de exacebação de desigualdades, de falta de transparência e de falhas de governança que é recorrente (CRAWFORD, 2021: 20-21).

Trata-se de mais uma dimensão a ser considerada pela empresa que adota soluções desse com esse perfil. Sua aplicação deve ser examinada dentro do contexto das políticas da empresa para governança, responsabilidade social e ambiental.

6. Considerações finais

Este texto, de caráter exploratório, buscou delimitar a aplicação do conceito de *Big Data* como o conjunto de técnicas e ambientes de processamento em que o volume de dados é grande ao ponto de demandar o processamento paralelo em múltiplos processadores para ser trabalhado. Apontou também aspectos típicos do tratamento desses

repositórios de dados, com um enfoque de mineração de dados, de identificação de aplicações inéditas e de análise preditiva.

A criação de novas informações e inferências, com um caráter estatístico, propicia intuições de grande valia, e consequente valor comercial e social, em aplicações tão diversas como pesquisas de informações na internet, identificação de preferências de clientes de estabelecimentos comerciais, direcionamento de propaganda político-eleitoral, condução de programas de saúde pública ou investigação criminal. Relataram-se alguns episódios, para ilustrar os resultados práticos oferecidos por essa abordagem.

A capacidade de processar grandes volumes de dados, enfrentando custos elevados, é um diferencial competitivo importante. Parte desses custos são fixos, pois decorrem da capacidade de manter uma bancada de equipamentos para armazenamento e tratamento de dados em grande escala e preservar um "lago de dados" a ser explorado. Nesses custos fixos pode-se incluir, também, a manutenção de uma estrutura organizacional e de equipes de profissionais habilitados no tratamento de *Big Data* e dotados de maturidade técnica para tal. E quem conduz um projeto, conduz vários.

Desse modo, empresas ou entidades com proficiência em *Big Data* dedicam-se a desenvolver projetos com escalas sempre crescentes, preservando um diferencial em relação aos competidores e, consequentemente, um posicionamento de imagem e de marca em seus mercados.

A proficiência no tratamento de *Big Data* depende de alguns fatores. O primeiro e mais primordial é o de que a técnica tem que fazer sentido para o negócio da empresa ou da entidade. Como foi apontado, *Big Data* envolve custos elevados e só faz sentido se o retorno for compensador. Soluções convencionais são, em muitos casos, mais apropriadas. Um segundo fator é o de que há um ciclo de aprendizado e maturação a ser considerado, que varia com a aplicação e com a disponibilidade de dados. Um terceiro fator é o de que, apesar de ser um mercado ainda em desenvolvimento, já dispõe de soluções técnicas experimentadas e de empresas de consultoria e suporte consolidadas internacionalmente e com intensa atuação no Brasil.

Em particular, a oferta de serviços de armazenamento de terceira parte (*data warehouse*) é um mercado que merece atenção. Na avaliação mais superficial, *data warehouses* são vistos como grandes instalações com extensas bancadas de equipamentos, destinadas a acumular enormes quantidades de dados. No entanto, o conceito envolve aspectos mais complexos. Os dados não são acumulados na forma em que são recebidos. Para que possam ser adequadamente recuperados, são organizados em ambientes mais simples e isolados, os *data marts*. Cada *data mart* envolve um grande bloco de dados que atendam a uma mesma estrutura ou característica, e vários conjuntos de atributos ou dimensões usados para fazer distintas formas de acesso ordenado a esses dados. Desse modo, a incorporação de dados a um *data warehouse* envolve passos de extração desses dados de outros repositórios, de organização desses dados, de correção de erros dos mesmos e de indexação adequada.

O caráter de construção de conhecimento ou de *insights* a respeito de pessoas ou comunidades traz ao debate o tratamento jurídico dessas informações, que embora probabilísticas, carregam um elevado grau de acerto e se referem, em muitos casos,

a aspectos de temperamento, preferências ou hábitos dos quais os próprios interessados não estão cientes.

Além disso, o debate sobre a propriedade dessas informações é controverso, pois em muitos casos são informações pessoais em sentido lato, mas criadas ou construídas por terceiros.

Há, portanto, aspectos em que o empresário, ao decidir pela adoção de *data analytics*, fica exposto a riscos, em termos de segurança jurídica. O referencial jurídico e regulatório está defasado em relação às perspectivas dessa tecnologia e, apesar de avanços recentes, como a LGPD, ainda há controvérsias éticas e jurídicas acerca do tratamento de dados pessoais e dos limites ao uso e à comercialização de informações, dos direitos sobre esses dados e suas relações, da garantia de continuidade no acesso a dados e da responsabilização sobre decisões automáticas.

O espaço para o debate acerca de *Big Data* é, portanto, amplo, e requer uma postura criteriosa, diante das perspectivas de mercado que se abrem. É preciso aceitar o fato de que essas técnicas, que vêm sendo usadas de modo mais disseminado há cerca de uma década, ainda deverão evoluir, ampliando os espaços de aplicação. Portanto, trata-se de uma tecnologia que, embora intensamente usada, está ainda em transformação.

Deve ser destacado, mais uma vez, que este texto tem pretensão meramente exploratória, de modo a abrir caminho a análises subsequentes. Não representa, portanto, um quadro descritivo do estado-da-arte do *Big Data*. Apreciações ulteriores fazem-se necessárias para que esse panorama seja completamente desenhado.

Referências bibliográficas

ADAMS, Nick, Ted NORDHAUS e Michael SHELLENBERGER (2011). Counterterrorism since 9/11: Evaluating the Efficacy of Controversial Tactics. Oakland (CA), EUA: Breakthrough Institute.

ANDRADE, Elvira, Paulo B. TIGRE, Lourença F. SILVA, Denise F. SILVA, Joaquim C. de MOURA, Rosângela V. de OLIVEIRA, Arlan SOUZA (2007). Propriedade Intelectual em Software: o que podemos apreender da experiência internacional?". *Revista Brasileira de Inovação – RBI*, 6 (1): 31-53.

ANDREU-PEREZ, Javier, Carmen Y POON, Robert D. MERRIFIELD, Stephen C. WONG e GUANG-ZHONG Yang (2015). "Big Data for health". *IEEE Journal of Biomedical and Health Informatics*, 19 (4): 1193-1208.

BACHRACH, Yoram, Michal KOSINSKI, Thore GRAEPEL, Pushmeet KOHLI e David STILLWELL (2012). "Personality and patterns of Facebook usage". ACM Web Science Conference. *Proceedings of the ACM Web Science Conference*, pp 36–44.

CIBEIRA, Fernando (2017). Macristocracia: la Historia de las Familias que Gobiernan la Argentina. Buenos Aires: Planeta.

CHUI, Michael, James MANYIKA, Jacques BUGHIN, Richard DOBBS, Charles ROXBURGH, Hugo SARRAZIN, Geoffrey SANDS e Magdalena WESTERGREN (2012). "The social economy: unlocking value and productivity through social technologies".

Relatório do McKinsey Global Institute, julho de 2012. Disponível em http://www.mckinsey.com/industries/high-tech/our-insights/.

COMPUTERWORLD (2017). "Mercado brasileiro de big data e analytics fatura US$ 1,16 bi e já representa quase 50% da AL". *Computerworld*, 21/3/2017. Disponível em: http://computerworld.com.br/.

CRAWFORD, Kate (2021). Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence. New Haven (CN), EUA: Yale.

DYCHÉ, Jill (2014). "Technology for Big Data". Em: DAVENPORT, Thomas A. Big Data at Work. Boston (MA), EUA: Harvard Business Review Press.

GODA, Kazuo e Masaru KITSUREGAWA (2012). "The history of storage systems". *Proceedings of the IEE*, 100: 1433-1440, maio de 2012.

GRASSEGER, Hannes e Mikael KROGERUS (2017). "A manipulação da democracia através do Big Data". *Jornal CGN*, 6 de fevereiro de 2017.

HILL, Kashmir (2012). "How Target figured out a teen girl was pregnant before her father did". *Forbes Tech*, 16/2/2012. Disponível em https://www.forbes.com/sites/kashmirhill/.

ISAACSON, Walter (2014). The Innovators: How a Group of Hackers, Geniuses, and Geeks Created the Digital Revolution. Nova York (NY), EUA: Simon&Schuster.

KOSINSKI Michal, David STILLWELL e Thore GRAEPEL (2013). "Private traits and attributes are predictable from digital records of human behavior". *Proceedings of the National Academy of Sciences – PNAS*, 110 (15): 5802-5805, abril de 2013.

LEMOS, Ronaldo e Massimo DI FELICE (2015). A Vida em Rede. Campinas: Papirus.

LIBORIO, Marcus (2017). "Lava Jato conta com 'Big Data' da corrupção, afirma Moscardi Grillo". *JC Net*, 26/4/2017. Disponível em: http://www.jcnet.com.br/Geral/2017/04/.

LINS, Bernardo E. (2013). "A evolução da internet: uma perspectiva histórica". *Cadernos Aslegis*, 17(48): 11-45.

LINS, Bernardo E. (2019). "Mídia digital e formação da preferência eleitoral". *Comunicação & Sociedade*, 41(1): 271-306.

MACHADO, Jorge S., Pablo ORTELLADO e Márcio M. RIBEIRO (2015). Xeque-Mate: o tripé da proteção de dados pessoais no jogo de xadrez das iniciativas legislativas no Brasil. São Paulo: GPoPAI/USP.

MANIYIKA, James, Michael CHUI, Brad BROWN, Jacques BUGHIN, Richard DOBBS, Charles ROXBURGH e Angela Hung BYERS (2011). "Big Data: the next frontier for innovatrion, competition, and productivity". Relatório do McKinsey Global Institute, maio de 2011. Disponível em http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/.

MAYER-SCHÖNBERGER, Viktor e Kenneth CUKIER (2013). Big Data: a Revolution that will Transform how we Live, Work, and Think. Boston (MA), EUA: Eamon Dolan/HMH.

McAFEE, Andrew e Erik BRYNJOLFSSON (2012). "Big Data: the management revolution". *Harvard Business Review*, 60-68, outubro de 2012.

McCUE, Colleen (2015). Data Mining and Predictive Analysis. Oxford, UK: Butterworth-Heinemann, 2ª ed.

MUNDIE, Craig (2014). "Privacy pragmatism: focus on data use, not data collection". *Foreign Affairs*, 93 (2): 28-38.

MURDOCH, Travis B. e Allan S. DETSKY (2013). "The inevitable application of Big Data to health care". *Journal of the American Medical Association (JAMA)*, 309 (13): 1351-1352.

RODRIGUES, Flávia C., João B. BERBERT e , Maria Luiza F. TEIXEIRA (2013). "Proteção intelectual para programas de computador: considerações acerca da possibilidade de patenteamento do software". *Revista de Direito Empresarial – RDEmp*, 10 (1): 205-220.

TIGRE, Paulo B. e Felipe S. MARQUES (2009). "Apropriação tecnológica na economia do conhecimento: inovação e propriedade intelectual de software na América Latina". *Economia e Sociedade*, 3 (37): 547-566.

VIEIRA, Marcos R., Josiel .M. FIGUEIREDO, Gustavo LIBERATTI, Alvaro M. VIEBRANTZ (2012). Bancos de Dados NoSQL: Conceitos, Ferramentas, Linguagens e Estudos de Casos no Contexto de Big Data. Minicurso. Simpósio Brasileiro de Bancos de Dados - SBBD 2012. Mimeo.