Capítulo 6

O QUE MUDA COM BIG DATA

"Não podemos prever o futuro, mas podemos criá-lo." — Peter Drucker

Diante de todas as características, requisitos e possibilidades de aplicações de Big Data, podemos perceber o quanto ele pode exercer um papel transformador nas mais diversas áreas de negócio. O grande volume de informações tem possibilitado a criação de novos produtos e serviços de dados, novas formas de se otimizar uma tarefa e novas informações para a tomada de decisão estratégica.

Mas para que esses benefícios possam ser alcançados, é necessário inovar, em inúmeros aspectos. Neste capítulo final, você confere aspectos sobre como Big Data está sendo um agente de transformação nas empresas.

6.1 CULTURA ORIENTADA POR DADOS

Espero que tenha conseguido demonstrar a você como os dados podem oferecer conhecimentos valiosos se forem bem explorados. Temos agora a oportunidade de capturar uma imensidão de informação e tornar isso um produto ou serviço.

Mas será que as empresas estão preparadas para essa cultura, na qual os dados passam a ter um papel chave dentro da organização?

Trabalhar com Big Data requer uma mudança não somente técnica; é preciso uma mudança de comportamento.

Seja em uma startup ou em uma empresa de milhares de funcionários, é preciso se preocupar em criar meios para que os dados sejam utilizados eficientemente. Mas o que deve ser feito para isso?

Não existe uma receita mágica para fazer com que se crie uma cultura orientada por dados dentro da empresa. Porém, existem algumas abordagens que podem facilitar o alcance desse objetivo. Um dos entraves para se criar essa cultura é a criação de silos de dados. Podemos definir isso como conjuntos de bancos de dados da empresa que não se comunicam com outros sistemas, sendo usados de forma isolada.

Tradicionalmente utilizada pelas grandes empresas, essa maneira de separar os dados — de forma que somente um pequeno grupo tenha acesso a um conjunto específico de informação — impede que os profissionais explorem todos os dados da empresa. Além disso, pelo fato de que diferentes conjuntos de dados nunca são integrados, perde-se aí a possibilidade de identificar problemas, ou responder a questões que só surgiriam com uma visão completa deles.

Por isso é importante que a empresa crie alternativas para que todos tenham acesso a todos os dados da empresa. Conforme ilustrado na figura a seguir, em busca de gerar essa democratização dos dados, algumas empresas estão adotando um conceito chamado Data Lake. Data Lake é um ambiente no qual os dados estruturados e não estruturados, coletados de diferentes fontes, são armazenados em uma única plataforma (por exemplo, um cluster com ecossistema Hadoop), permitindo a integração dos dados e a geração de análises por meio de tecnologias de Big Data.

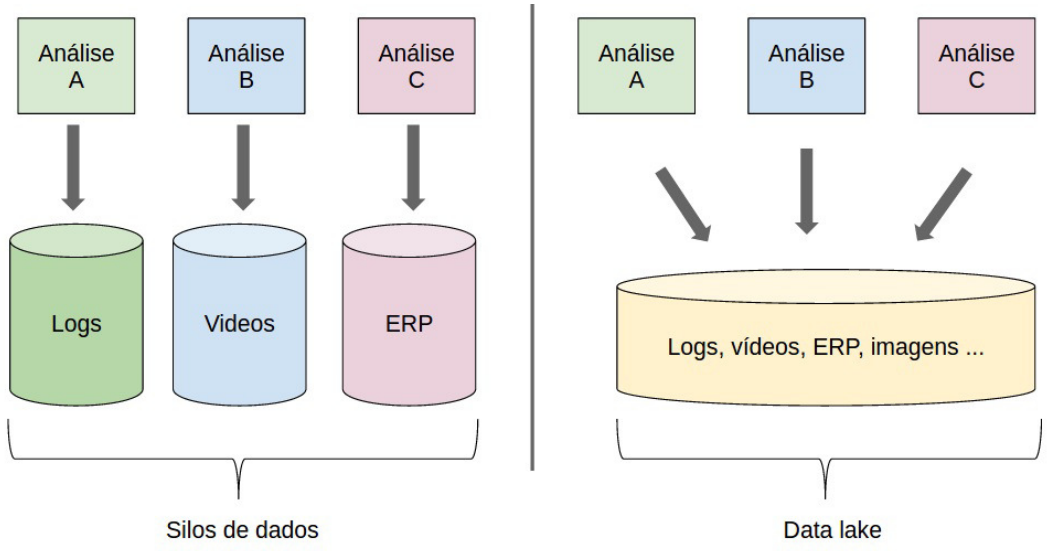

Figura 6.1: Diferença entre silos de dados e data lake

DJ Patil, atualmente *Chief Data Officer* da Casa Branca dos Estados Unidos, mencionou em seu livro *Building Data Science Teams* que uma organização preparada para ter uma cultura orientada por dados é aquela que "*adquire, processa e utiliza dados em tempo hábil para criar eficiências, iterar e desenvolver novos produtos e navegar no cenário competitivo*".

Por esse motivo, é importante também que a organização tenha a iniciativa de coletar e explorar diferentes fontes de dados. Isso possibilita a descoberta de dados relevantes para a empresa. Certamente muitos dados serão descartados, porém, com essa abordagem, a empresa poderá descobrir novas fontes de conhecimento.

Outra abordagem importante para incentivar a cultura orientada por dados é permitir que os profissionais criem e testem suas ideias. Mais importante do que uma prova de conceito, é permitir que esses profissionais consigam criar uma prova de valor sobre sua ideia, demonstrando os benefícios obtidos da solução proposta. A partir de prova de valor, a proposta pode ser aperfeiçoada dentro da empresa, para então entrar em seu portfólio

de soluções.

Sua empresa, ou a empresa na qual você trabalha, tem esse comportamento? Caso a resposta seja não, saiba que ainda são poucas as que também possuem essa maturidade em culturas orientadas por dados.

Entretanto, é importante que exista um movimento em relação a isso, para que os profissionais tenham o ambiente ideal para atuar em projetos de Big Data. Com certeza isso não é garantia de sucesso nos projetos de Big Data, porém a falta desse comportamento pode inviabilizar a execução dos projetos.

6.2 A CARREIRA DO CIENTISTA DE DADOS

Juntamente com a popularidade de Big Data, a profissão cientista de dados também se tornou notória. Entretanto, assim como Big Data, não há ainda uma definição específica do que vem a ser esse profissional e quais conhecimentos são necessários para se tornar um.

Esse termo foi cunhado por DJ Patil, como um meio de classificar uma equipe de profissionais que trabalhavam com os produtos e serviços de dados, que segundo Patil, trabalhavam com os dados e com a ciência para criar novas soluções. Por isso o nome cientista de dados.

Conforme as empresas foram se interessando por Big Data e deram início a provas de conceito sobre esse tema, elas passaram a buscar esse profissional. O problema é que elas exigiam (algumas ainda exigem) que o cientista de dados a ser contratado tivesse conhecimentos profundos em ciência da computação, programação, matemática, banco de dados, estatística, design e negócios.

O que se costuma dizer é que empresas assim não estão em

busca de um profissional, e sim de um unicórnio. Pois sabemos que não é possível (ou no mínimo muito raro) existir um profissional com tamanha habilidade. Por isso, fique calmo, você não precisará saber tudo isso para atuar em um projeto de Big Data.

As empresas atualmente estão mais conscientes de que um projeto de Big Data deve ser formado por uma equipe heterogênea, com profissionais de diferentes habilidades trabalhando em conjunto para o desenvolvimento da solução. Com isso, foram surgindo carreiras específicas para cada especialidade, como o analista de dados, o engenheiro de dados, o gerente de projetos de Big Data, o administrador de infraestrutura de Big Data, entre outros.

Por exemplo, vimos no decorrer do livro 4 estágios distintos em um projeto de Big Data: captura e armazenamento dos dados, processamento dos dados, análise dos dados e visualização dos dados. Se você tem interesse em atuar especificamente em alguma dessas áreas, veja a seguir sugestões de áreas em que você deveria se especializar:

- **Captura e armazenamento dos dados:** banco de dados relacional, banco de dados NoSQL, linguagem SQL, frameworks para transferência de dados em lote, estrutura e gerenciamento de dados, armazenamento de dados na nuvem, frameworks para transferência de dados em streaming, sistemas de arquivos distribuídos, técnicas de agregação de dados, armazenamento de dados em data warehouse, armazenamento de dados em data lakes, técnicas de tolerância a falhas.
- **Processamento de dados:** frameworks de processamento distribuído em lote e em tempo real, processamento de dados na nuvem, linguagens de programação, computação em nuvem, testes,

otimização, técnicas de escalabilidade, disponibilidade, administração de cluster, latência e desempenho de aplicações.

- **Análise de dados:** técnicas de mineração de dados, técnicas de aprendizado de máquina, métodos estatísticos, técnicas de filtragem e limpeza dos dados, expressão regular, matrizes, modelagem, frameworks e bibliotecas para manipulação de dados, frameworks de análise de dados em ambientes distribuídos.

- **Visualização de dados:** conhecimento em experiência do usuário, design, computação gráfica, *storytelling*, programação Web, frameworks para visualização de dados, habilidades em comunicação.

Além de ter profissionais qualificados para cada um desses estágios, é importante em um projeto de Big Data que esses profissionais tenham a capacidade de conversar entre eles e com os profissionais da empresa que possuem conhecimento do negócio em que o projeto será aplicado. Para isso, é necessário que esses profissionais tenham, ao mesmo tempo, conhecimento profundo sobre a área que atuam, mas também tenham um conhecimento geral sobre as outras áreas.

Esse tipo de profissional é também conhecido como o profissional "Tipo T", conforme apresentado na figura seguinte. Ou seja, mesmo que você seja um especialista no processamento de dados em Hadoop, por exemplo, é importante que você tenha conhecimentos básicos sobre outras áreas, como coleta e visualização de dados.

Essa abordagem evita situações como a de um gestor do projeto solicitar algo para os analistas que seja impossível de ser modelado, ou que os analistas preparem os dados de uma maneira que

inviabilize a construção de gráficos dinâmicos.

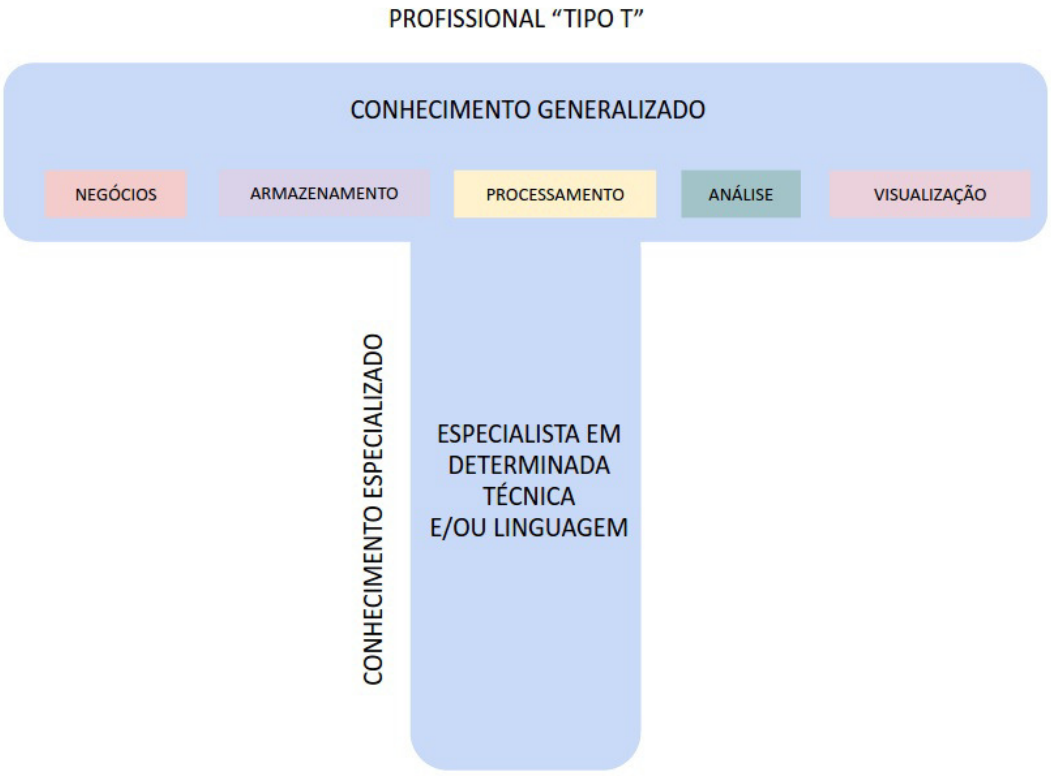

Figura 6.2: Profissional tipo T

Além dessas habilidades, a comunicação entre todos os profissionais durante todo o projeto é essencial. Por esse motivo, ser comunicativo é uma característica tão esperada dos cientistas de dados. Além do conhecimento técnico e habilidades em comunicação, listo a seguir outras características esperadas em um cientista de dados:

- **Curiosidade** — Em eventos de Big Data e Ciência de Dados que participei nos últimos anos, a curiosidade é citada como uma das principais características de um cientista de dados. Ele deve ter curiosidade e disposição para aprender sobre o negócio em que a empresa atua, sobre as diferentes áreas dentro da organização, e sobre diferentes tecnologias e métodos de análises de dados.

- **Colaboração** — Essa característica está aliada à comunicação. Uma equipe de cientista de dados deve ter em mente que eles devem colaborar entre si e com os demais profissionais da empresa, para alavancar a cultura orientada por dados.

- **Criatividade** — Ter um conhecimento técnico sobre a área que você atua é de suma importância para trabalhar com Big Data. Entretanto, isso não é suficiente. Trabalhar com dados exige que você exerça seu lado criativo, pensando em formas inovadoras de se utilizar os dados.

- **Pensamento analítico** — É importante que um cientista de dados saiba fazer perguntas sobre os dados, bem como tenha a capacidade de analisar os resultados e entender o que eles expressam. São esses profissionais que poderão auxiliar a empresa para obter valor sobre os dados.

- **Comprometimento** — Trabalhar com uma vasta quantia de dados de fontes internas e externas é algo complexo, e exige paciência e prática para chegar ao resultado final com sucesso. Para isso, é preciso que cada profissional tenha o comprometimento de fazer o seu melhor na área em que possui expertise.

Concluindo, você não poderá ser um cientista de dados se não estiver disposto a aprender (continuamente) novas tecnologias, novos meios de resolver problemas. Nem se não tiver interesse em trabalhar em equipe com profissionais de outras áreas, e em descobrir o que os dados podem revelar. Porém, se você acredita que se enquadra nesse perfil de cientista de dados, saiba que há muitas empresas procurando por você.

Mas onde você pode estudar para ter essas habilidades técnicas necessárias para atuar com Big Data? Isso ainda é um problema quando falamos da carreira dos cientistas de dados.

No momento ainda há poucas instituições que oferecem a capacitação para essa profissão, sendo que as que oferecem fazem isso há poucos anos, portanto, não possuem muitos alunos formados até o momento. Como resultado, temos uma oferta maior do que a demanda. Atualmente, a procura por profissionais capacitados é um dos grandes desafios das empresas que atuam ou desejam atuar com Big Data.

Além dos cursos oferecidos pelas instituições, existem outras alternativas que podem alavancar o conhecimento em Big Data. É possível, por exemplo, utilizar bancos de dados abertos disponíveis em diversos sites da Web para começar a explorá-los e assim aprender mais sobre esse processo.

Caso não saiba o que fazer com esses dados, uma alternativa é usar sites que lhe ofereçam não somente os dados, mas desafios na utilização deles. Um desses sites é o Kaggle (https://www.kaggle.com/), uma plataforma para cientistas de dados onde empresas lançam desafios, dos quais usuários e equipes de usuários podem competir em busca de resolvê-los. Além disso, muitas empresas já utilizam esse portal para encontrar profissionais que apresentem habilidades em Big Data.

Outra alternativa para aperfeiçoar o conhecimento em Big Data é a participação em hackathons. Sendo uma combinação das palavras em inglês *hack* e *marathon*, esse nome é dado para competições com foco na prototipação de uma solução tecnológica em um curto período de tempo (de 24 a 48 horas, normalmente).

Mesmo não sendo restrito a Big Data, é comum que em muitas dessas competições sejam criados desafios que envolvem a

manipulação de grande volume de dados. Além de ter a oportunidade de desenvolver um produto que desperte o interesse de investidores, nesses eventos você tem a oportunidade de conhecer diversas pessoas (inclusive profissionais renomados) que atuam na mesma área que a sua, aumentando o seu networking.

6.3 A PRIVACIDADE DOS DADOS

Embora o cientista de dados seja considerado atualmente uma das carreiras mais promissoras quando se fala em Big Data, além dele, existe um outro profissional que certamente atuará em larga escala na era do Big Data: o advogado especialista em violação de privacidade de dados.

O aumento do volume de dados gerados por pessoas e dispositivos ocorreu de forma acelerada. Juntamente com esses dados, surgiram milhares de aplicações com diferentes serviços orientados a dados. Enquadram-se aqui exemplos como os sistemas de recomendações utilizados por sites de e-commerce, streaming de músicas e vídeos, a publicidade personalizada e a análise de sentimento em redes sociais.

Esses e inúmeros outros serviços surgiram de forma tão avassaladora, que não foi possível identificar previamente os limites e o impacto do uso de tais serviços. Como resultado, muitos deles estão coletando um grande número de informações dos usuários, como localização, hábitos de compra, estado de saúde, hábitos de leitura, o que come, o que veste, o que assiste, como se exercita. Esse cenário desencadeia uma série de possíveis problemas de privacidade.

Um dos grandes problemas com esses serviços é que, na maioria dos casos, os dados são capturados sem a anuência do usuário. O usuário utiliza o serviço sem ter conhecimento de que, por exemplo,

os dados do seu carro estão sendo coletados pela seguradora, que as informações de sua localização geográfica estão sendo usadas para oferecer promoções, ou que o modo como ele navega nas páginas Web também está sendo utilizado para identificar padrões de comportamento na internet.

Já temos exemplos reais de problemas relacionados à privacidade de dados. Talvez um dos mais notáveis seja o ocorrido com a empresa Target. Essa grande empresa americana no setor varejista usou estatísticas de compras para criar um modelo capaz de prever quais mulheres estavam grávidas, para assim fazer a propaganda de produtos direcionadas a elas.

Isso foi feito avaliando comportamentos nos históricos de compras que estivessem relacionadas à gravidez, como por exemplo, a compra de loções e suplementos de cálcio e zinco. Esse modelo possibilitou a Target a enviar cupons de desconto para essas mulheres.

O problema ocorreu quando a empresa recebeu uma reclamação de um homem, solicitando uma explicação sobre o fato de sua filha adolescente ter recebido cupons para compras de roupas de bebês. Entretanto, alguns dias depois, o mesmo homem voltou a falar com a empresa dizendo que a filha assumiu que, de fato, estava grávida.

Esse problema dá origem a diversas questões: como garantir que os dados capturados e utilizados para traçar seu perfil estão de fato condizentes? Como ter a opção de escolher quais informações podem ser usadas por terceiros? Como criar uma política de privacidade que me mantenha protegido no uso dos dados de terceiros?

Outra questão deve ser avaliada quando falamos em Big Data: a possível discriminação com base na análise dos dados. Uma empresa que decide não recrutar um usuário para um emprego

devido ao seu histórico de comportamento nos sites e nas redes sociais está agindo de forma ética? E se ela recusar ocorrer ao usuário que desejar obter um cartão de crédito, um seguro de vida ou uma vaga em uma universidade?

Sabemos que discriminação é algo ilegal, mas como identificar que ela ocorre no contexto de Big Data? Como saber que dados ilícitos estão sendo utilizados? Temos aqui um grande problema a ser debatido, principalmente para evitar que Big Data tenha um impacto negativo nas classes mais vulneráveis.

Infelizmente, a pessoa ou empresa que sofrer alguma violação de privacidade ainda possui pouco respaldo da legislação. Essa situação ocorre em nível global.

Cada governo está se adaptando para regulamentar questões de violação de privacidade. No Brasil, tivemos a iniciativa do Marco Civil da Internet, contendo um conjunto de regras e guias sobre a utilização dos dados pelos serviços. Entretanto, essa tarefa é muito complexa e difícil de ser controlada.

Big Data é um conceito poderoso e pode oferecer inúmeros benefícios. Para evitar que sua utilização gere problemas de violação de privacidade, muitas questões ainda precisam ser debatidas e novos procedimentos devem ser criados, tais como o maior controle individual sobre os dados coletados pelas empresas, maior transparência no uso desses dados e maior segurança em relação ao armazenamento, utilização e divulgação dos dados.

6.4 NOVOS MODELOS DE NEGÓCIOS

Estamos vivendo em uma era de transição, em que soluções disruptivas estão sendo criadas, transformando negócios e mudando o status quo de como as coisas funcionam. Já temos diversos

exemplos de soluções que estão tendo sucesso nesse sentido.

Um desses exemplos é o Uber, uma aplicação com foco no transporte de passageiros. Essa empresa trouxe um novo modelo de negócios para esse segmento, que até então era realizado somente por taxistas.

Com sua proposta colaborativa, outras pessoas que possuem um carro passam a oferecer o mesmo serviço, podendo assim cobrar um valor do serviço mais barato, já que eles não pagam as mesmas taxas que os taxistas. Por sua característica disruptiva, a aceitação desse novo serviço não é bem vista por todos, tanto que estamos acompanhando como está sendo polêmica a entrada dessa solução no Brasil.

Outras duas empresas que estão revolucionando o segmento em que atuam é o AirBnB e o Netflix. O AirBnB (*AirBed & Breakfast* — cama de ar e café da manhã) trouxe uma nova alternativa às pessoas que precisam de uma hospedagem temporária, além dos tradicionais serviços oferecidos pelos hotéis.

Com essa solução, tornou-se possível alugar um quarto, casa ou apartamento de milhares de pessoas do mundo todo, que oferecem seu espaço para alugar dentro do aplicativo AirBnB. O resultado foi uma nova forma das pessoas terem lucro com seu imóvel, além de uma infinidade de alternativas de hospedagem. Com certeza essa solução impactou os negócios do ramo hoteleiro.

O mesmo impacto está causando a Netflix, na indústria da televisão. Esse serviço trouxe uma nova forma de assistir filmes e seriados, na qual o usuário paga uma assinatura mensal e tem a sua disponibilidade filmes e seriados para que ele possa assistir no dia e horário que quiser.

Além disso, nos últimos anos, a empresa passou a criar

conteúdo exclusivo de filmes e séries, atraindo ainda mais os usuários para adquirirem o serviço. Imaginem quantas pessoas estão deixando de assistir programas da TV aberta, ou estão cancelando sua TV por assinatura, para utilizar somente o serviço da Netflix. Com certeza ela está revolucionando a forma com que vamos assistir televisão daqui para a frente.

Além dessas 3 soluções, podemos identificar outros serviços que mudaram a experiência dos usuários. Por exemplo, qual é o método que você atualmente mais ouve músicas? A maioria das pessoas aderiram aos serviços online de streaming, como o Spotify, Pandora e Deezer.

Se pensarmos que a 10 anos atrás não utilizávamos o smartphone para ouvir música e muito menos serviços como esses, podemos perceber como em pouco tempo houve uma transformação na indústria musical. Nesse estilo inovador, temos também serviços como o Dropbox para compartilhamento de arquivos, e o Coursera para aulas online.

Mas imagino que você esteja pensando: qual a relação dessas soluções com Big Data? Pois bem, sem tecnologias de Big Data para oferecer uma infraestrutura escalável e de alta disponibilidade para essas soluções, possivelmente não seria possível criar uma solução colaborativa que fosse utilizada por milhões de usuários 24 horas por dia.

Além disso, essas soluções usam os dados de forma criativa para conseguir analisar e extrair informações que tornem seus serviços ainda mais eficientes. Por exemplo, a própria Netflix captura dados do comportamento do usuário, tais como: quais filmes foram assistidos, quais os usuários desistiram no meio, qual cena o usuário voltou a assistir, em qual momento o usuário adiantou um filme e quais foram as categorias de filme mais assistidas.

Todas essas métricas são usadas para que o serviço seja oferecido de forma cada vez mais personalizada ao usuário, e também serve de insights para a escolha do conteúdo na produção das séries e filmes. Ou seja, Big Data tem um papel crucial nessas soluções, pois somado às capacidades oferecidas pelos avanços da Internet, da computação móvel e de computação em nuvem, elas estão abrindo portas para a criação de soluções que rompem as barreiras existentes nas soluções tradicionais.

O que é importante notar é que essa mudança que estamos vivenciando está apenas em seu início. Ainda há muito que pode ser criado utilizando as tecnologias que temos disponíveis atualmente. Por isso a inovação é tão importante quando falamos em Big Data.

Além de todas as questões técnicas envolvidas em um projeto, é preciso criatividade para extrair o melhor que Big Data pode oferecer. Você trabalha na área de agricultura, publicidade, telecomunicação, saúde, esporte, biologia, manufatura, direito, ou qualquer outra área? Você pode pensar em como Big Data pode auxiliar o negócio em que você atua.

Por ser uma abordagem ainda recente, você tem grandes possibilidades de se destacar com sua solução proposta. Eu sei, não é uma tarefa fácil, mas sua oportunidade é agora.

6.5 MENSAGEM FINAL

Chegamos ao fim do último capítulo do livro. Espero que você tenha se interessado por essa jornada ao mundo de Big Data e que tenha se motivado a atuar nessa área tão promissora.

O que mais me motiva em Big Data é saber que temos a possibilidade de utilizar algo que temos em abundância (o grande volume de dados) e que, com isso, podemos criar soluções

inovadoras. Saber que você pode gerar valor por organizar os dados que antes não eram compreendidos, por contar uma história que esclareça o porquê dos fatos ocorridos, por disponibilizar informações de forma tão rápida que permita acelerar a resolução dos problemas, por criar novos produtos e serviços que mudem a vida das pessoas para melhor. Isso não é sensacional?

Se este livro foi um dos primeiros contatos que você teve com Big Data, você deve estar ciente de que a jornada para se aprofundar no tema é longa. Entretanto, ela pode ser muito divertida e cheia de descobertas. O que pode acontecer durante o caminho é você se apaixonar por dados e criar uma relação eterna com eles.

Bem-vindo ao mundo de Big Data!

Para saber mais

1. BARLOW, Mike. *Learning to Love Data Science.* O'Reilly Media, Inc., 2015.
2. BARLOW, Mike. *The culture of big data.* O'Reilly Media, Inc., 2013.
3. HARRIS, Harlan; VAISMAN, Marck; MURPHY, Sean Gordon. *Analyzing the Analyzers:* An Introspective Survey of Data Scientists and Their Work. O'Reilly Media, Inc., 2013.
4. KING, John; MAGOULAS, Roger. *Data science salary survey:* tools, trends, what pays (and what doesn't) for data professionals. O'Reilly Media, Inc., 2016. Disponível em: http://www.oreilly.com/data/free/2016-data-science-salary-survey.csp.
5. PATIL, DJ; MANSON, Hilary. *Data Driven:* Creating a Data Culture. O'Reilly Media, Inc., 2015. Disponível em: http://www.oreilly.com/data/free/data-driven.csp.