Sumário Resumido

Prefácio

Vivemos em tempos empolgantes, até mesmo revolucionários. Quando nossas interações diárias passam do mundo físico para o digital, quase toda atitude que tomamos gera dados. As informações transbordam de nossos dispositivos móveis e de cada troca online. Os sensores e máquinas coletam, armazenam e processam informações sobre o ambiente à nossa volta. Conjuntos de dados novos e enormes agora são abertos e acessíveis ao público.

Essa inundação de informações nos dá o poder de tomar decisões mais conscientes, reagir mais rapidamente à mudança e compreender melhor o mundo que nos rodeia. Contudo, pode ser trabalhoso saber por onde começar para compreender todo esse dilúvio de dados. Quais dados devem ser coletados? Quais métodos devem ser considerados? E, o mais importante, como obtemos respostas a partir dos dados para responder às perguntas mais urgentes sobre nossos negócios, vidas e mundo?

O data science é o segredo para tornar essa tempestade de informações útil. Simplificando, ele é a arte de coletar dados para prever nosso comportamento futuro, descobrir padrões que nos ajudarão a priorizar ou obter informações produtivas ou, ainda, extrair sentidos desse montante de dados inexplorados.

Com frequência, digo que uma das minhas versões favoritas da palavra "big", em big data, é "expansivo". A revolução dos dados se espalha por tantas áreas que agora é obrigatório que todos os profissionais, independentemente de sua área de atuação, compreendam como usá-los, exatamente como as pessoas tiveram que aprender a manejar computadores nas décadas de 1980 e 1990. Este livro é planejado para ajudá-lo a fazer isso.

Tenho visto pessoalmente o quanto o conhecimento de data science pode transformar radicalmente as organizações e o mundo. Na DataKind, utilizamos o poder do data science a serviço da humanidade, engajando especialistas na área e de setores sociais para trabalhar em projetos que lidam com os problemas críticos humanitários. Também orientamos discussões sobre como o data science pode ser aplicado para resolver os maiores desafios do mundo. Desde usar imagens de satélite e estimar os níveis de pobreza, até problematizar décadas de violações dos direitos humanos e impedir mais atrocidades, as equipes na DataKind têm trabalhado com muitas organizações humanitárias diferentes e sem fins lucrativos apenas para iniciar suas jornadas nesse campo. Uma lição ressoa através de cada projeto que fazemos: as pessoas e organizações mais comprometidas em usar dados de maneiras novas e responsáveis são as que mais terão sucesso nesse novo cenário.

Apenas segurar este livro já representa que você também está dando seus primeiros passos nessa jornada. Se você é um pesquisador experiente procurando

atualizar algumas técnicas de data science ou é completamente novato no mundo dos dados, o *Data Science Para Leigos* irá prepará-lo com as habilidades necessárias para vislumbrar o que pode realizar. Você conseguirá expor novas descobertas a partir de seus dados materiais, desde apresentar novas informações e a última campanha de marketing até compartilhar aprendizados sobre como prevenir a propagação de doenças.

Estamos de fato em uma era de vanguarda, e aqueles que aprenderem o data science conseguirão fazer parte dessa nova aventura emocionante, guiando nosso caminho por onde passarmos. Para você, essa aventura começa agora. Bem-vindo a bordo!

Jake Porway

Fundador e diretor-executivo da DataKind

Introdução

O poder do big data e do data science está revolucionando o mundo. Das empresas modernas às escolhas de estilo de vida do cidadão digital de hoje, as informações do data science conduzem mudanças e melhorias em áreas diversas. Embora o data science seja um assunto desconhecido para muitos, qualquer pessoa que deseje fazer a diferença em sua carreira precisa compreendê-lo.

Este livro é um manual de referência para orientá-lo nas complexas áreas englobadas por big data e data science. Se você deseja entender o que acontece no mundo ao seu redor, este é o livro. Se for um gerente organizacional que busca compreender como as implementações de data science e big data podem melhorar seu negócio, este é o livro. Se for um analista técnico, ou mesmo um desenvolvedor, que anseia por um livro de consulta rápida sobre como a aprendizagem de máquina e os métodos de programação funcionam no universo do data science, este é o livro.

Mas se estiver buscando um treinamento prático aprofundado em áreas muito específicas envolvidas em implementar, de fato, as iniciativas de data science e big data, este *não* será o melhor livro para você. Procure em outro lugar, porque esta obra é um manual breve e geral sobre *todas* as áreas englobadas por data science e big data. Para manter o livro para leigos, não me aprofundo ou restrinjo a nenhuma área. Muitos cursos online estão disponíveis para dar suporte às pessoas que desejam despender tempo e energia explorando essas frestas estreitas. Sugiro que as pessoas complementem este material com cursos em áreas específicas de seu interesse.

Embora outras obras que tratam deste tema tendam a se concentrar no uso do Microsoft Excel para aprender as técnicas básicas de data science, *Data Science Para Leigos* vai além, apresentando a linguagem de programação estatística R, Python, D3.js, SQL, Excel e muitos aplicativos de fonte aberta que você pode usar para começar a praticá-lo. Alguns livros sobre data science são desnecessariamente prolixos, com os autores girando em círculos e sem chegar ao ponto. Não aqui. Diferentemente dos livros academicistas, muito conservadores, escrevi este livro em uma linguagem amistosa e acessível — porque o data science é um assunto amistoso e acessível!

Para ser honesta, até agora o reino do data science tem sido dominado por alguns gênios selecionados que tendem a apresentar o assunto de maneira desnecessariamente técnica e intimidadora. O data science básico não é tão confuso nem difícil de entender. O *data science* é simplesmente a prática de usar um conjunto de técnicas e metodologias analíticas para derivar e comunicar informações

úteis e valiosas a partir de dados brutos. A finalidade do data science é otimizar os processos e dar suporte à tomada de decisão melhorada e com dados relatados, gerando, assim, um aumento no valor — com o *valor* sendo representado por várias vidas salvas, dólares preservados ou porcentagem de rendimentos aumentada. Em *Data Science Para Leigos* apresento muitos conceitos e abordagens que você poderá usar ao extrair informações valiosas de seus dados.

Muitas vezes, os cientistas de dados ficam tão envolvidos analisando a casca das árvores que simplesmente esquecem de procurar o caminho para sair da floresta. Essa armadilha comum é a que você deve evitar a todo custo. Trabalhei muito para assegurar que este livro apresente a finalidade essencial de cada técnica de data science, e os objetivos que você pode conseguir utilizando-as.

Sobre Este Livro

De acordo com o padrão *Para Leigos*, este livro se organiza em um formato modular e fácil de acessar, que lhe permite que o use como um guia útil e uma referência prática. Em outras palavras, você não precisa lê-lo por completo, do início ao fim. Basta ir na parte que deseja e relegar o resto. Tomei muito cuidado para usar exemplos reais que ilustram conceitos de data science que podem ser muito abstratos. Os endereços da web e códigos de programação aparecem em monofonte.

Penso que...

Ao escrever este livro, supus que os leitores estão minimamente aptos do ponto de vista técnico para dominar as tarefas avançadas no Microsoft Excel — tabelas dinâmicas, agrupamento, classificação, plotagem e outros. Ter grandes habilidades em álgebra, estatística básica ou até em cálculo comercial ajuda também. Bobagem ou não, espero muito que todos os leitores tenham uma especialização em algum tema no qual possam aplicar o que apresento neste livro. Como os cientistas de dados devem ser capazes de entender intuitivamente as implicações e aplicações das informações derivadas de dados, o especialista no assunto é o componente-chave do data science.

Ícones Usados Neste Livro

Quando você avançar neste livro, verá os seguintes ícones nas margens:


DICA

Este ícone marca as dicas (sério?) e os atalhos que você pode seguir para facilitar o domínio do assunto.

LEMBRE-SE

Este ícone registra informações especialmente importantes a saber. Para obtê-las, basta ver o material apresentado por esses ícones.

PAPO DE ESPECIALISTA

Este ícone destaca informações de natureza altamente técnica, que você normalmente pode pular.

CUIDADO

Este ícone aconselha a ter cautela! Ele sinaliza as informações importantes, que evitam dores de cabeça.

Além Deste Livro

Este livro inclui os seguintes recursos externos:

- **Folha de Cola do Data Science:** Este livro vem com uma Folha de Cola prática, que lista os atalhos úteis, assim como as definições resumidas dos processos essenciais e conceitos descritos no livro. Você pode usá-la como uma referência rápida e fácil ao fazer o data science. Para obter a Folha de Cola, basta acessar `www.altabooks.com.br` e buscar por *Folha de Cola Data Science* na caixa Pesquisa.
- **Conjuntos de Dados para Tutorial de Data Science**: Este livro tem alguns tutoriais que contam com conjuntos de dados externos. Você pode baixá-los no repositório GitHub deste curso em `https://github.com/BigDataGal/Data-Science-for-Dummies` (conteúdo em inglês), também disponível para download em `www.altabooks.com.br` (procure pelo título do livro).

De Lá para Cá, Daqui para Lá

Apenas para enfatizar, a estrutura articulada deste livro permite que você escolha e comece a ler em qualquer lugar desejado. Embora não precise ler do início ao fim, alguns bons capítulos para iniciar são os Capítulos 1, 2 e 9.

1 Começando

NESTA PARTE . . .

Conheça o data science.

Defina o big data.

Explore soluções para os problemas do big data.

Veja como os negócios reais fazem bom uso do data science.

NESTE CAPÍTULO

- Usando o data science em diferentes setores
- Reunindo os diferentes componentes do data science
- Identificando as soluções viáveis do data science para seus próprios desafios de dados
- Dominando o mercado com o data science

Capítulo **1**

Compreendendo o Data Science

Há algum tempo, *todos* nós somos completamente inundados por dados. Eles vêm de cada computador, dispositivo móvel, câmera e sensor imaginável — e agora, até de relógios e tecnologias em roupas. Os dados são gerados em toda interação de mídia social que fazemos, todo arquivo que salvamos, foto que fazemos e pesquisa que realizamos; são até gerados quando fazemos algo tão simples quanto pedir ao mecanismo de busca favorito instruções para chegar à sorveteria mais próxima.

Embora a imersão em dados não seja nova, é notável que o fenômeno está se acelerando. Lagos, poças e rios de informação transformaram-se em inundações e verdadeiros tsunamis de dados estruturados, semiestruturados e não estruturados, que jorram de quase toda atividade ocorrida nos mundos digital e físico. Bem-vindo ao universo do *big data!*

Se você for como eu, pode ter imaginado: "O que interessa em todos esses dados? Por que usar recursos complexos para gerá-los e coletá-los?". Apesar de, até uma década atrás, ninguém estar bem preparado para usar grande parte dos dados gerados, as tendências hoje mudaram definitivamente. Especialistas conhecidos como *engenheiros de dados* constantemente encontram maneiras inovadoras e poderosas de capturar, combinar e condensar volumes enormes de

dados de modos inimagináveis, e outros especialistas, conhecidos como *cientistas de dados*, tirando informações úteis e valiosas a partir desses dados.

Na verdade, o data science representa a otimização de processos e recursos. Ele produz *informações de dados* — conclusões ou previsões úteis derivadas de dados que você pode usar para entender e melhorar seu negócio, investimentos, saúde e até seu estilo de vida e vida social. Usar as informações do data science é como conseguir enxergar no escuro. Para qualquer objetivo ou busca imaginada, é possível encontrar métodos do data science para ajudar a prever a rota mais direta de onde você está até onde deseja estar — e antecipar cada desvio na estrada entre os dois lugares.

Vendo Quem Pode Usar o Data Science

Os termos *data science* e *engenharia de dados* geralmente são mal usados e confusos, portanto, começarei esclarecendo que esses dois campos são, na verdade, domínios de especialização separados e distintos. *Data science* é a ciência da computação que extrai informações significativas a partir de dados brutos e comunica-as, com eficiência, em atividades práticas. A *engenharia de dados*, por outro lado, é um domínio da engenharia dedicado a criar e manter sistemas que superam as obstruções do processamento de dados e os problemas do tratamento de dados para os aplicativos que consomem, processam e armazenam grandes volumes, variedades e velocidades de dados. No data science e na engenharia de dados, normalmente você trabalha com estas três variedades de dados:

- **Estruturados:** Os dados são armazenados, processados e manipulados em um sistema de gerenciamento do banco de dados relacional (RDBMS).
- **Não estruturados:** Os dados são normalmente gerados a partir de atividades humanas e não se encaixam em um formato de banco de dados estruturado.
- **Semiestruturados:** Os dados não se encaixam em um sistema de banco de dados estruturado, todavia, são estruturados por etiquetas (tags) úteis para criar uma forma de ordem e hierarquia nos dados.

Muitas pessoas acreditam que apenas as grandes organizações, com bons recursos financeiros, implementam as metodologias do data science para otimizar seu negócio, mas este não é o caso. A proliferação de dados criou uma demanda por informações, e essa demanda está incorporada em muitos aspectos de nossa cultura moderna — desde o passageiro de Uber que espera o motorista para pegá-lo exatamente na hora e local previstos pelo aplicativo, até o visitante online que espera que a plataforma Amazon recomende as melhores alternativas de produto para poder comparar com outros, análogos, antes de fazer uma compra. Os dados e a necessidade de informações baseadas neles são onipresentes. Como todos os tipos de organização reconhecem que estão

mergulhados em um ambiente competitivo, do tipo cada um por si e com base em dados, a habilidade com os dados surge como uma função essencial e indispensável em praticamente toda linha de negócio.

O que isso significa para a pessoa comum? Primeiro, que os funcionários devem cada vez mais dar suporte a um conjunto de exigências tecnológicas que avança progressivamente. Por quê? Bem, quase todas as indústrias confiam cada vez mais nas tecnologias de dados e nas informações que incitam. Como consequência, muitas pessoas têm uma necessidade contínua de aprimorar suas habilidades técnicas, ou enfrentam a possibilidade real de serem substituídas por um funcionário com mais conhecimento tecnológico.

A boa notícia é que atualizar essas habilidades técnicas geralmente não requer que as pessoas voltem para a faculdade — Deus me livre —, sejam graduadas em estatística, ciência da computação ou de dados. A má notícia é que, mesmo para um profissional treinado e autodidata, sempre é um trabalho extra manter-se relevante no setor e ter conhecimento técnico. Quanto a isso, a revolução dos dados não é tão diferente de nenhuma outra mudança que atingiu a indústria no passado. O fato é que, para fazer a diferença, você precisa de tempo e esforço para adquirir apenas as habilidades que o mantêm atualizado. Quando estiver aprendendo o data science, você poderá fazer alguns cursos, aprender sozinho com recursos online, ler livros sobre o tema e ir a eventos em que poderá aprender o que precisa saber para permanecer no jogo.

Quem pode usar o data science? Você. Sua organização. Seu patrão. Qualquer pessoa que tenha um pouco de compreensão e treinamento pode usar as informações dos dados para melhorar suas vidas, carreiras e a prosperidade de seus negócios. O data science representa uma mudança no modo como você aborda o mundo. Quanto aos resultados exatos, as pessoas geralmente costumavam fazer adivinhações, agir e esperar pelo resultado desejado. Porém, com as informações dos dados, agora elas têm acesso a uma visão preditiva necessária para realmente orientar a mudança e conseguir os resultados de que precisam.

Você pode usar as informações dos dados para fazer mudanças nas seguintes áreas:

- **Sistemas comerciais:** Otimize o retorno em investimento (o ROI fundamental) para qualquer atividade mensurável.
- **Desenvolvimento técnico da estratégia de marketing:** Use as informações de dados e a análise preditiva para identificar as estratégias que funcionam, elimine os esforços abaixo do desempenho e teste novas abordagens.
- **Mantenha as comunidades seguras:** Aplicativos de policiamento preditivo ajudam as pessoas que aplicam a lei a prever e antecipar atividades locais criminosas.
- **Torne o mundo um lugar melhor para os menos afortunados:** Os cientistas de dados das nações desenvolvidas usam dados sociais, móveis e de sites para gerar análises reais que melhoram a eficiência da resposta humanitária a desastres, epidemias, problemas de escassez de alimentos e outros.

Analisando as Peças do Quebra-cabeça

Para usar o data science de forma prática, no verdadeiro significado do termo, você precisa de um conhecimento analítico de matemática e estatística, habilidades de codificação necessárias para trabalhar com os dados e uma mínima especialização no assunto. Sem ela, você pode dizer-se apenas matemático ou estatístico. Igualmente, um programador de software sem especialização no assunto e sem prática analítica pode ser considerado mais um engenheiro de software ou desenvolvedor, mas não um cientista de dados.

Como a demanda por informações provenientes dos dados aumenta exponencialmente, toda área é forçada a adotar o data science. Assim, diferentes tipos de data science surgiram. A seguir, estão apenas alguns títulos sob os quais os especialistas de toda disciplina usam o data science: cientista de dados consultor tecnológico (ad tech), diretor de análise digital bancária, cientista de dados clínicos, cientista de dados da engenharia geográfica, cientista de dados de análise geoespacial, analista político, cientista de dados de personalização varejista e analista de informática clínica em farmacometria. Visto que sem um planejamento estratégico aparentemente ninguém controla quem é um cientista de dados, nas seções a seguir explico os principais componentes que fazem parte de qualquer função do data science.

Coletando, consultando e consumindo dados

Os engenheiros de dados têm o trabalho de capturar e combinar grandes volumes de *big data* estruturado, não estruturado e semiestruturado — os dados que excedem a capacidade de processamento dos sistemas de banco de dados convencionais porque são grandes demais, movem-se muito rápido ou não se encaixam nos requisitos estruturais da arquitetura de banco de dados tradicional. Novamente, as tarefas da engenharia de dados são separadas do trabalho realizado no data science, que se concentra mais em análise, previsão e visualização. Apesar dessa distinção, sempre que os cientistas de dados coletam, consultam e consomem dados durante o processo de análise, realizam um trabalho parecido com o do engenheiro de dados (a função anteriormente exposta neste capítulo).

Embora informações valiosas possam ser geradas a partir de uma única fonte de dados, em geral, a combinação de várias fontes relevantes fornece as informações contextuais necessárias para orientar melhor as decisões baseadas em dados. Um cientista de dados pode trabalhar com os vários conjuntos de dados armazenados em um único banco de dados ou até com armazenamentos de dados diferentes. (Para saber mais sobre como combinar os conjuntos de dados, veja o Capítulo 3.) Em outras ocasiões, os dados de origem são armazenados e processados em uma plataforma em nuvem, criada por engenheiros de software e de dados.

Não importa como os dados são combinados ou onde são armazenados, se você for um cientista de dados, quase sempre terá que *consultá-los*, ou seja, escrever comandos para extrair os conjuntos de dados relevantes a partir dos sistemas de armazenamento. Na maioria das vezes, você usará o Structured Query Language (SQL) para consultá-los. (O Capítulo 16 é sobre SQL, portanto, se a abreviação o assusta, vá para o capítulo agora.)

Se você usa um aplicativo ou faz uma análise personalizada com uma linguagem de programação como R ou Python, pode escolher entre vários formatos de arquivo aceitos universalmente:

- **Arquivos com valores separados por vírgula (CSV):** Quase toda marca de desktop e aplicativo de análise da web aceita esse tipo de arquivo, assim como as linguagens de script usadas comumente, como Python e R.
- **Scripts:** A maioria dos cientistas de dados sabe usar a linguagem de programação Python ou R para analisar e visualizar os dados. Esses arquivos de script terminam com a extensão `.py`, `.ipynb (Python)` ou `.r (R)`.
- **Arquivos do aplicativo:** O Excel é útil para fazer análises de amostragem, rápidas e fáceis, em conjuntos de dados pequenos e médios. Esses arquivos de aplicativo têm a extensão `.xls` ou `.xlsx`. Os aplicativos de análise geoespacial, como ArcGIS e QGIS, salvam com seus próprios formatos de arquivo patenteados (extensão `.mxd` para o ArcGIS e extensão `.qgs` para o QGIS).
- **Arquivos de programação da web:** Se você estiver criando visualizações de dados baseadas na web e personalizadas, poderá trabalhar no D3.js — ou no Data-Driven Documents, uma biblioteca JavaScript para a visualização de dados. Quando você trabalha no D3.js, usa dados para manipular os documentos da web utilizando os arquivos `.html`, `.svg` e `.css`.

Aplicando a modelagem matemática nas tarefas do data science

O data science conta muito com as habilidades matemáticas de um profissional (na área estatística, como descrito na seção a seguir) precisamente porque são necessárias para interpretar os dados. Essas habilidades também são valiosas no data science porque você pode usá-las para fazer previsões, modelagem de decisão e teste de hipótese.


LEMBRE-SE

A *matemática* usa métodos determinísticos para formar uma descrição *quantitativa* (ou *numérica*) do mundo; a *estatística* é uma forma de ciência derivada da matemática, mas foca em usar uma abordagem *estocástica* (de probabilidades) e métodos inferenciais para elaborar uma descrição quantitativa do mundo. Mais sobre isso é analisado no Capítulo 5.

Os cientistas de dados usam métodos matemáticos para criar modelos de decisão, gerar aproximações e fazer previsões sobre o futuro. O Capítulo 5 apresenta muitas abordagens matemáticas aplicadas complexas que são úteis ao trabalhar com data science.


LEMBRE-SE

Para fins deste livro, suponho que você tenha um conjunto de habilidades bem sólido em matemática básica — seria bom se tivesse habilitação em cálculo ou até em álgebra linear. Porém, tento alcançar os leitores no nível em que estão. Entendo que você pode trabalhar com base em um conhecimento matemático limitado (álgebra avançada ou, talvez, cálculo comercial), portanto, passo os conceitos matemáticos avançados usando uma abordagem simples e fácil, para todos entenderem.

Derivando informações de métodos estatísticos

Em data science, os métodos estatísticos são úteis para compreender melhor o significado de seus dados, validar hipóteses, simular cenários e fazer previsões de eventos futuros. Conhecimentos estatísticos avançados são um pouco raros, mesmo entre analistas quantitativos, engenheiros e cientistas; porém, se você quiser se dar bem em data science, reserve um tempo para entender alguns métodos estatísticos básicos, como regressão linear e logística, classificação naïve Bayes e análise da série temporal. Esses métodos estão no Capítulo 5.

Codificar, codificar, codificar — é apenas uma parte do jogo

A codificação é inevitável quando trabalhamos com data science. Você precisa saber escrever o código para instruir o computador a manipular, analisar e visualizar seus dados como deseja. As linguagens de programação, como Python e R, são importantes para escrever scripts para a manipulação, análise e visualização dos dados, e o SQL é útil para consultá-los. A biblioteca D3.js do JavaScript é uma nova opção para fazer visualizações de dados legais, personalizadas, interativas e baseadas na web.

Embora a codificação seja uma exigência do data science, ela não precisa ser a *coisa* assustadora que as pessoas falam por aí. Sua codificação pode ser tão extravagante e complexa quanto você deseja, mas também é possível adotar uma abordagem bem simples. Embora essas habilidades sejam primordiais para o sucesso, você pode aprender com muita facilidade a codificação básica para praticar um data science de alto nível. Dediquei os Capítulos 10, 14, 15 e 16 a ajudá-lo a entender como usar o D3.js para a visualização de dados da web, codificar no Python e R, e consultar no SQL (respectivamente).

Aplicando o data science em uma área de conhecimento

Os estatísticos têm mostrado certa obstinação ao aceitar a importância do data science. Muitos gritaram: "O data science não é novo! É apenas outro nome para o que fazemos desde sempre". Embora eu possa simpatizar com sua perspectiva, sou forçada a ficar com o grupo dos cientistas de dados que declara notoriamente que o data science é autônomo e definitivamente distinto das abordagens estatísticas que o compõem.

Minha postura sobre a natureza única do data science é baseada, até certo ponto, no fato de que os cientistas de dados geralmente usam as linguagens de computação não utilizadas na estatística tradicional e adotam abordagens derivadas do campo da matemática. Mas a principal distinção entre a estatística e o data science é a necessidade de especialização no assunto.

Como os estatísticos geralmente têm apenas uma especialização limitada nos campos fora da estatística, quase sempre são forçados a consultar um especialista para verificar exatamente o que suas descobertas significam e decidir a melhor direção na qual prosseguir. Por outro lado, os cientistas de dados precisam ter uma grande especialização na área na qual trabalham. Eles geram conceitos complexos e usam sua especialização de domínio para entender exatamente como esses conceitos se relacionam com a área em que atuam.

Esta lista descreve algumas maneiras como os especialistas no assunto aplicam o data science para melhorar o desempenho de seus respectivos setores:

- Os **engenheiros** usam a aprendizagem de máquina para otimizar a eficiência da energia no design de estruturação moderno.
- Os **cientistas de dados clínicos** trabalham na personalização dos planos de tratamento e usam a informática na assistência médica para prever e prevenir futuros problemas de saúde nos pacientes em risco.
- Os **cientistas dos dados de marketing** usam a regressão logística para prever e prevenir a *rotatividade de clientes* (a perda ou a oscilação de clientes de um produto ou serviço para um concorrente). Falo mais sobre como diminuir a rotatividade de clientes nos Capítulos 3 e 20.
- Os **jornalistas de dados** `limpam` os sites (em outras palavras, extraem dados em massa diretamente das páginas de um site) para obter dados atualizados para descobrir e informar os acontecimentos mais recentes. (Falo mais sobre o jornalismo de dados no Capítulo 18.)
- Os **cientistas de dados na análise de crimes** usam a modelagem preditiva espacial para prever, prevenir e impedir atividades criminosas. (Veja o Capítulo 21 para obter todos os detalhes sobre como usar o data science para descrever e prever a atividade criminosa.)
- As **pessoas que fazem o bem com o uso de dados** usam a aprendizagem de máquina para classificar e dar informações vitais sobre as comunidades

afetadas por desastres, para dar um suporte de decisão em tempo real na resposta humanitária, que você pode ler no Capítulo 19.

Comunicando ideias com dados

Como cientista de dados, você deve ter habilidades de comunicação oral e escrita aguçadas. Se um cientista de dados não puder comunicar-se, todo o conhecimento e informação no mundo não servirão de *nada* para sua organização. Os cientistas de dados precisam conseguir explicar as informações de dados de um modo que os membros da equipe possam entender. Não apenas isso, precisam conseguir produzir visualizações de dados e narrativas escritas claras e significativas. Na maioria das vezes, as pessoas precisam ver algo por si mesmas para entender. Os cientistas de dados devem ser criativos e pragmáticos em seus meios e métodos de comunicação. (Trato dos tópicos da visualização de dados e narrativa baseada em dados com mais detalhes nos Capítulos 9 e 18, respectivamente.)

Explorando as Alternativas de Solução do Data Science

As organizações e seus líderes ainda lutam com o melhor uso de big data e data science. A maioria sabe que a análise avançada é capaz de conferir uma tremenda vantagem competitiva para suas organizações, mas poucos têm ideia das opções disponíveis ou benefícios exatos que o data science oferece. Nesta seção, apresento três alternativas principais de solução do data science e descrevo os benefícios que sua implementação acarreta.

Montando a própria equipe interna

Muitas organizações acham que é financeiramente vantajoso estabelecer a própria equipe interna de profissionais dedicados aos dados. Isso economiza o dinheiro que gastariam para conseguir resultados parecidos contratando consultores independentes ou implantando uma solução de análise já pronta baseada em nuvem. Três opções para criar uma equipe data science interna são:

- **Treinar os funcionários existentes.** Se você quiser equipar sua organização com o poder do data science e análise, o treinamento em data science (alternativa mais barata) poderá tornar o pessoal existente muito especializado, com capacitação em dados, para compor sua equipe interna.
- **Contratar pessoas treinadas.** Algumas organizações atendem a seus requisitos contratando cientistas de dados experientes ou recém-formados em data science. O problema dessa opção é que não há pessoas suficientes por aí, e se você quiser encontrar pessoas que queiram participar,

comumente exigirão altos salários. Lembre-se, além das exigências de matemática, estatística e codificação, os cientistas de dados devem ter um alto nível de especialização no campo específico no qual atuam. Por isso é extremamente difícil encontrá-los. Até que as universidades tornem a habilitação em dados uma parte integrante de todo programa educativo, encontrar cientistas de dados altamente especializados e capacitados para atender às necessidades organizacionais será praticamente impossível.

- **Treinar os funcionários existentes e contratar alguns especialistas.** Outra boa opção é treinar os funcionários existentes para as tarefas de data science de alto nível e, então, trazer alguns cientistas de dados experientes para atender às exigências de estratégia e solução de problemas mais avançadas do data science.

Terceirizando as exigências para os consultores particulares de data science

Muitas organizações preferem terceirizar suas necessidades de data science e análise para um especialista externo, usando uma das duas estratégias gerais:

- **Abrangente:** Esta opção atende à organização inteira. Para criar uma implementação avançada de data science para sua empresa, você pode contratar um consultor particular para ajudar no desenvolvimento de uma estratégia abrangente. Esse tipo de serviço provavelmente terá um custo, mas você recebe informações muitíssimo valiosas em retorno. Um estrategista conhecerá as opções disponíveis para atender a seus requisitos, assim como os benefícios e desvantagens de cada uma. Com a estratégia em mãos e um especialista de plantão disponível para ajudá-lo, você poderá gerir com muita facilidade a tarefa de criar uma equipe interna.
- **Individual:** É possível aplicar soluções fragmentadas em problemas específicos que surgem, ou surgiram, em sua organização. Se você não estiver preparado para o processo complexo do design abrangente de estratégia e implementação, poderá contratar trabalhos específicos de um consultor particular de data science. Essa abordagem de tratamento pontual ainda oferece vantagens de data science sem precisar que você reorganize a estrutura e as finanças da organização inteira.

Aproveitando as soluções de plataforma em nuvem

Uma solução em nuvem possibilita a análise de dados para profissionais que têm apenas um nível modesto de capacitação. Alguns viram a explosão do big data e data science vindo de longe. Embora ainda seja recente para a maioria, os profissionais e organizações bem informados vêm trabalhando rápido e com empenho

para se preparar. Aplicativos em nuvem novos e privados, como o Trusted Analytics Platform ou TAP (`http://trustedanalytics.org`) (conteúdo em inglês), se dedicam a tornar mais fácil e rápido que as organizações implementem suas iniciativas de big data. Outros serviços em nuvem, como o Tableau, oferecem opções automatizadas de dados e código de fonte aberta — desde a modelagem estatística simples e básica até a análise e a visualização de dados. Embora você ainda precise entender a relevância estatística, matemática e substancial das informações de dados, aplicativos como o Tableau oferecem excelentes resultados sem precisar que os usuários saibam escrever código ou scripts.


LEMBRE-SE

Se você decidir usar soluções de plataforma baseadas em nuvem para ajudar sua organização a atingir seus objetivos com o data science, ainda precisará de uma equipe interna treinada e habilitada para projetar, executar e interpretar os resultados quantitativos dessas plataformas. A plataforma não eliminará a necessidade de um treinamento interno e qualificação em data science; ela simplesmente potencializará sua organização para que atinja mais prontamente seus objetivos.

Permitindo que o Data Science o Torne Mais Competitivo

Neste livro, espero mostrar a força do data science e como você pode usá-la para atingir mais rapidamente seus objetivos pessoais e profissionais. Não importa o setor no qual você trabalha, adquirir habilidades em data science poderá transformá-lo em um profissional com maior representação no mercado. A lista a seguir descreve apenas alguns dos principais setores que podem aproveitar o data science e a análise:

- **Corporações, empresas de pequeno e médio portes (SMEs) e empresas e-commerce:** Otimização dos custos de produção, maximização das vendas, aumentos do ROI de marketing, otimização da produtividade da equipe, redução da rotatividade de clientes, aumento de valor no ciclo de vida do cliente, exigências de inventário e previsão de vendas, otimização do modelo de preços, detecção de fraudes, filtro de colaboração, mecanismos de recomendação e melhorias na logística.
- **Governos:** Otimização dos processos de negócio e produtividade da equipe, melhorias no suporte de decisão do gerenciamento, previsão de finanças e orçamento, controle e otimização dos gastos e detecção de fraudes.
- **Academia:** Melhorias na alocação de recursos e no gerenciamento do desempenho dos alunos, reduções de abandonos, otimização dos processos de negócio, previsão de finanças e orçamento, e aumento do ROI de recrutamento.

NESTE CAPÍTULO

- Definindo big data
- Vendo algumas fontes de big data
- Diferenciando data science e engenharia de dados
- Reforçando no Hadoop
- Explorando soluções para os problemas de big data
- Verificando um projeto de engenharia de dados real

Capítulo **2**

Explorando Encadeamentos e Infraestrutura da Engenharia de Dados

Há muita euforia em torno do big data atualmente, mas a maioria das pessoas realmente não sabe nem entende o que é ou como pode usá-lo para melhorar suas vidas e profissões. Este capítulo define o termo big data, explica de onde vem e como é usado, e descreve as funções que os engenheiros e cientistas de dados desempenham no ambiente do big data. Neste capítulo, apresento os conceitos fundamentais do big data, dos quais você precisa para começar a gerar suas próprias ideias e planos sobre como aproveitar o big data e o data science para melhorar seu estilo de vida e fluxo de negócio. (***Sugestão:*** você conseguiria melhorar seu estilo de vida dominando algumas das tecnologias analisadas neste capítulo — o que certamente levaria a mais oportunidades para chegar a uma posição bem remunerada, que também oferece excelentes benefícios para o estilo de vida.)

Definindo Big Data com Três Vs

Big data são os dados que excedem a capacidade de processamento dos sistemas de banco de dados convencionais porque são muito grandes, movem-se muito rápido ou não se encaixam nos requisitos estruturais das arquiteturas de bancos de dados comuns. Se os volumes de dados se classificam em terabytes ou petabytes, as soluções da engenharia de dados devem ser planejadas para atender aos requisitos de destino e uso pretendidos dos dados.


PAPO DE ESPECIALISTA

Quando se fala sobre dados comuns, provavelmente você ouve as palavras *kilobyte* e *gigabyte* usadas como medidas — 10^3 e 10^9 bytes, respectivamente. Por outro lado, quando se fala em big data, palavras como *terabyte* e *petabyte* vêm à tona — 10^{12} e 10^{15} bytes, respectivamente. Um *byte* é uma unidade de dados com 8 bits.

Três características (conhecidas como "os três Vs") definem o big data: volume, velocidade e variedade. Como os três Vs do big data estão se expandindo frequentemente, tecnologias de dados mais recentes e inovadoras devem ser continuamente desenvolvidas para gerenciar os problemas do big data.


LEMBRE-SE

Em uma situação na qual é necessário adotar uma solução de big data para resolver um problema causado pela velocidade, volume ou variedade de seus dados, você ultrapassou o domínio dos dados comuns, e o problema que tem em mãos é de big data.

Lutando com o volume de dados

O limite inferior do volume do big data começa em apenas 1 terabyte e não há um limite superior. Se sua organização possui, pelo menos, 1 terabyte de dados, provavelmente é uma boa candidata a uma implementação do big data.


CUIDADO

Em sua forma bruta, grande parte do big data tem um *valor baixo*, ou seja, a proporção entre valor e dados é baixa no big data bruto. O big data é composto de números enormes de transações muito pequenas com vários formatos. Esses componentes adicionais de big data produzem um valor real apenas depois de serem agregados e analisados. Os engenheiros de dados têm o trabalho de prepará-los e os cientistas de dados, de analisá-los.

Lidando com a velocidade dos dados

Muito big data é criado com processos automatizados e instrumentação atualmente e, como seus custos de armazenamento são relativamente baixos, a velocidade do sistema é, muitas vezes, o fator limitador. O big data tem baixo valor. Como consequência, você precisa de sistemas capazes de consumir muitos dados em curto prazo para gerar informações rápidas e valiosas.

Nos termos da engenharia, *velocidade dos dados* é o volume de dados por tempo de unidade. O big data entra em um sistema médio em velocidades que variam de 30 kilobytes (K) por segundo até 30 *gigabytes* (GB) por segundo. Muitos sistemas com engenharia de dados precisam ter uma latência inferior a 100 milissegundos, medida desde o momento em que os dados são criados até quando o sistema responde. As exigências da velocidade de processamento podem facilmente chegar a mil mensagens por segundo nos sistemas de big data! Os dados que se movem em tempo real e com alta velocidade são um obstáculo para tomar decisões rápidas. A capacidade das tecnologias para o tratamento e o processamento dos dados geralmente limita as velocidades deles.


LEMBRE-SE

Há vários tipos de ferramentas de consumo de dados. Algumas das mais populares são descritas nesta lista:

- **Apache Sqoop:** Você pode usar essa ferramenta de transferência de dados para transferi-los rapidamente entre um sistema de dados relacional e o *sistema de arquivos distribuídos Hadoop (HDFS)* — ele usa clusters de servidores comuns para armazenar o big data. O HDFS possibilita, financeiramente, o tratamento e armazenamento de big data ao distribuir tarefas de armazenamento nos clusters de servidores comuns e baratos. É o sistema de armazenamento principal usado na implementação do big data.
- **Apache Kafka:** Esse sistema de mensagens distribuído atua como um agente de mensagens pelo qual as mensagens podem ser enviadas, e obtidas, para o HDFS. É possível usar o Kafka para consolidar e facilitar as chamadas de dados e os envios que os consumidores fazem para e a partir do HDFS.
- **Apache Flume:** Esse sistema distribuído basicamente lida com os dados de registro e eventos. Você pode usá-lo para transferir grandes quantidades de dados não estruturados para e a partir do HDFS.

Lidando com a variedade de dados

O big data fica ainda mais complexo quando você adiciona dados não estruturados e semiestruturados às fontes de dados estruturadas. Esses dados com *alta variedade* vêm de muitas fontes. O ponto principal é que são compostos de uma combinação de conjuntos de dados com arquétipos específicos subjacentes (estruturados, não estruturados ou semiestruturados). Os dados heterogêneos e com alta variedade são geralmente compostos por qualquer combinação de dados gráficos, arquivos JSON, arquivos XML, dados de mídia social, dados tabulares estruturados, dados de blogues e dados gerados a partir do fluxo de cliques.

Os dados *estruturados* podem ser armazenados, processados e manipulados em um sistema de gerenciamento do banco de dados relacional (RDBMS) tradicional. Esses dados podem ser gerados por pessoas ou máquinas, e são derivados de todos os tipos de fontes, desde fluxos de cliques e formulários da web até transações de ponto de venda e sensores. Os dados *não estruturados* são completamente autônomos — comumente gerados a partir de atividades humanas e não se encaixam em um formato de banco de dados estruturado. Tais dados podem ser derivados de postagens de blogues, e-mails e documentos do Word. Os dados *semiestruturados* não se encaixam em um sistema de banco de dados estruturado, mas são estruturados por tags úteis para criar ordem e hierarquia nos dados. Os semiestruturados são comumente encontrados em bancos de dados e sistemas de arquivos. Eles podem ser armazenados como arquivos de log, arquivos XML ou arquivos de dados JSON.


DICA

Fique familiarizado com o termo *data lake* — ele é utilizado pelos profissionais no setor de big data para se referir a um sistema de armazenamento de dados não hierarquizados que é usado para manter volumes enormes de dados multiestruturados em uma arquitetura plana de armazenamento. O HDFS pode ser usado como um repositório de armazenamento data lake, mas também é possível usar a plataforma Amazon Web Services S3 para atender aos mesmos requisitos em nuvem (a plataforma Amazon Web Services S3 é uma arquitetura de nuvem disponível para armazenar big data).

Identificando as Fontes de Big Data

O big data é continuamente gerado por pessoas, máquinas e sensores em todo lugar. As fontes típicas incluem dados de mídia social, transações financeiras, prontuários, fluxos de cliques, arquivos de log e a *Internet das Coisas* — uma rede de conexões digitais que reúne o conjunto sempre em expansão de dispositivos eletrônicos que usamos em nosso cotidiano. A Figura 2-1 mostra várias fontes populares de big data.

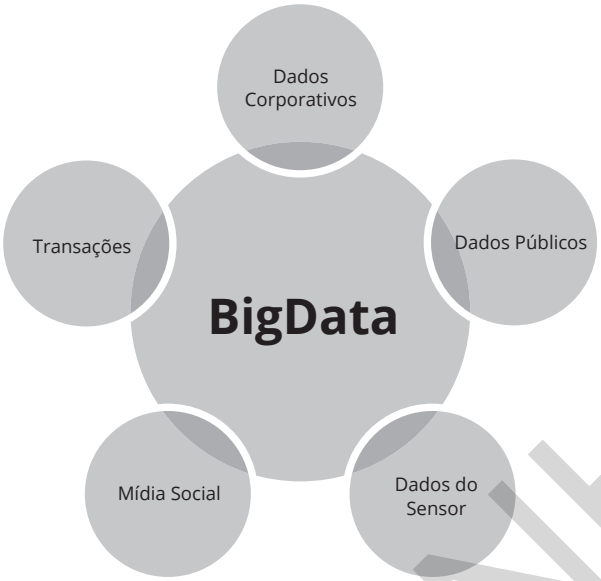

FIGURA 2–1: Fontes populares de big data.

Entendendo a Diferença entre Data Science e Engenharia de Dados

Data science e engenharia de dados são dois ramos diferentes no *paradigma do big data* — uma abordagem na qual velocidades, variedades e volumes enormes de dados estruturados, não estruturados e semiestruturados são capturados, processados, armazenados e analisados usando um conjunto de técnicas e tecnologias completamente novas em comparação com as usadas décadas atrás.

São úteis para derivar conhecimento e informações úteis de dados brutos. São elementos essenciais para qualquer sistema abrangente de suporte de decisão, e necessários para formular estratégias robustas para um futuro gerenciamento e crescimento do negócio. Embora os termos *data science* e *engenharia de dados* geralmente sejam usados alternadamente, são domínios de especialização distintos. Nas seções a seguir, apresento os conceitos fundamentais para o data science e a engenharia de dados e, então, mostro as diferenças de como funcionam no sistema de processamento de dados de uma organização.

Definindo data science

Se a *ciência* é um método sistemático com o qual as pessoas estudam e explicam fenômenos específicos do domínio que ocorrem na natureza, você pode

considerar o *data science* como o domínio científico dedicado à descoberta de conhecimento via análise de dados.


PAPO DE ESPECIALISTA

O termo *específico do domínio* refere-se ao setor ou campo cujos métodos de data science são usados para explorar.

Os cientistas de dados usam técnicas matemáticas e abordagens algorítmicas para derivar soluções para problemas corporativos e científicos complexos. Os profissionais do data science usam métodos preditivos para derivar informações que seriam impossíveis de outro modo. Nos negócios e na ciência, os métodos do data science auxiliam nas tomadas de decisão mais robustas:

- **Nos negócios,** a finalidade do data science é fornecer às organizações informações de dados necessárias para otimizar seus processos burocráticos e ter uma máxima eficiência e geração de renda.
- **Na ciência,** os métodos do data science são usados para fornecer resultados e desenvolver protocolos para atingir os objetivos específicos em questão.

O data science é um campo amplo e multidisciplinar. Para ser considerado um verdadeiro cientista de dados, você precisa ter especialização em matemática e estatística, programação de computador e em seu próprio campo de domínio.

Usando suas habilidades de data science, é possível:

- Usar a aprendizagem de máquina para otimizar os usos da energia e reduzir as pegadas de carbono corporativas.
- Otimizar estratégias para conseguir os objetivos nos negócios e na ciência.
- Prever níveis de contaminação desconhecidos a partir de conjuntos de dados ambientais esparsos.
- Planejar sistemas automatizados de prevenção contra roubo e fraude para detectar anomalias e disparar alarmes com base em resultados algorítmicos.
- Construir mecanismos de recomendação de sites para usar nas aquisições de terras e desenvolvimento de bens imobiliários.
- Implementar e interpretar a análise preditiva e técnicas de previsão para a ampliação do valor de negócio resultante.

Os cientistas de dados devem ter uma especialização quantitativa extensa e diversificada para resolver esses tipos de problemas.


PAPO DE ESPECIALISTA

Aprendizagem de máquina é a prática de aplicar algoritmos para aprender, e fazer previsões automatizadas, com os dados.

Definindo a engenharia de dados

Se a *engenharia* é a prática de usar a ciência e a tecnologia para planejar e criar sistemas que resolvem problemas, é possível considerar a *engenharia de dados* como o domínio da engenharia dedicado a criar e manter sistemas de dados para superar as obstruções no processo de dados e problemas em seu tratamento, que surgem devido ao alto volume, velocidade e variedade do big data.

Os engenheiros de dados usam as habilidades da ciência da computação e engenharia de software para criar sistemas para resolver problemas, lidar e manipular os grandes conjuntos de dados. Os engenheiros, geralmente, têm experiência em trabalhar e criar estruturas de processamento em tempo real e plataformas de processamento paralelo em massa (MPP) (analisado posteriormente neste capítulo), assim como RDBMSs. Normalmente, eles codificam em Java, C++, Scala e Python. Sabem como implantar o Hadoop MapReduce ou Spark para lidar, processar e aprimorar o big data em conjuntos de dados com tamanho mais gerenciável. Para simplificar, em relação ao data science, a finalidade da engenharia de dados é planejar soluções de big data criando plataformas de processamento de dados coerentes, modulares e dimensionáveis a partir das quais os cientistas derivam informações.

LEMBRE-SE

A maioria dos sistemas planejados é *construída*; eles são desenvolvidos ou elaborados no mundo físico. Entretanto, a engenharia de dados é diferente. Envolve planejar, criar e implementar soluções de software para os problemas no mundo dos dados, que pode parecer abstrato quando comparado com a realidade física da Ponte Golden Gate ou da Represa de Aswan.

Usando a engenharia de dados, você pode, por exemplo:

- Criar aplicativos de Software como Serviço (SaaS) em grande escala.
- Criar e personalizar os aplicativos Hadoop e MapReduce.
- Planejar e criar bancos de dados relacionais e arquiteturas distribuídas em alta escala para o processo de big data.
- Criar uma plataforma integrada para resolver simultaneamente problemas de consumo de dados, armazenamento, aprendizagem de máquina e gerenciamento do sistema, tudo em uma só interface.

Os engenheiros de dados devem dispor de habilidades sólidas em ciência da computação, design de bancos de dados e engenharia de softwares para realizar esse tipo de trabalho.

PAPO DE ESPECIALISTA

Software como Serviço (SaaS) é um termo que descreve os serviços de software hospedados em nuvem que são disponibilizados para os usuários via internet.

Comparando cientistas e engenheiros de dados

Com frequência, as funções do cientista de dados e do engenheiro de dados são confundidas e entrelaçadas por gerentes de contratação. A maioria das descrições de cargos das empresas contratantes geralmente diverge quanto aos títulos e funções, ou simplesmente exige que os candidatos sejam da área de data science e engenharia de dados.


DICA

Ao contratar alguém para ajudá-lo a entender seus dados, defina os requisitos claramente antes de descrever a função. Como os cientistas de dados devem ser especializados nas áreas específicas em que atuam, esse requisito geralmente afasta a necessidade de os cientistas também terem especialização em engenharia de dados (embora alguns tenham experiência em plataformas de dados de engenharia). E se você contratar um engenheiro de dados com habilidades em data science, normalmente essa pessoa não terá muita especialização fora dessa área. Esteja preparado para chamar um especialista no assunto para ajudá-lo.

Como muitas organizações combinam e confundem as funções em seus projetos de dados, os cientistas de dados algumas vezes passam muito tempo aprendendo a fazer o trabalho de um engenheiro de dados e vice-versa. Para obter um produto de alta qualidade em menos tempo, contrate um engenheiro para processar seus dados e um cientista para entendê-los.

Por fim, lembre-se que engenheiro e cientista de dados são apenas duas pequenas funções em uma estrutura organizacional maior. Os gerentes, funcionários intermediários e líderes organizacionais também desempenham um papel significativo no sucesso de qualquer iniciativa baseada em dados. O principal benefício da incorporação do data science e da engenharia de dados aos seus projetos é a utilização de dados externos e internos para dar suporte às decisões da sua organização.

Entendendo os Dados no Hadoop

Como os três Vs do big data (volume, velocidade e variedade) não permitem lidar com o big data usando sistemas de gerenciamento de banco de dados relacionais tradicionais, os engenheiros tiveram que inovar. Para resolver as limitações dos sistemas relacionais, eles se voltaram para a plataforma de processamento de dados Hadoop e reduziram o big data a conjuntos de dados menores, mais gerenciáveis e passíveis de análise pelos cientistas.


LEMBRE-SE

Quando você ouvir as pessoas usarem o termo *Hadoop* atualmente, em geral estarão se referindo a um ecossistema Hadoop que inclui HDFS (para o armazenamento de dados), MapReduce (para o processamento de dados em massa), Spark (para o processamento de dados em tempo real) e YARN (para o gerenciamento de recursos).