1 Ciência de dados em um mundo de big data

Este capítulo cobre

- Definindo ciência de dados e big data
- Reconhecendo os diferentes tipos de dados
- Obtendo insights sobre o processo de ciência de dados
- Apresentando os campos da ciência de dados e big data
- Trabalhando com exemplos de Hadoop

Big data é um termo genérico para qualquer coleção de conjuntos de dados tão grandes ou complexos que se torna difícil processá-los usando técnicas tradicionais de gerenciamento de dados, como, por exemplo, RDBMS (sistemas de gerenciamento de banco de dados relacional). O RDBMS amplamente adotado tem sido considerado há muito tempo como uma solução única para todos, mas as demandas de manipulação de big data mostraram o contrário. *A ciência de dados* envolve o uso de métodos para analisar grandes quantidades de dados e extrair o conhecimento que eles contêm. Você pode pensar na relação entre big data e ciência de dados como sendo semelhante à relação entre o petróleo bruto e uma refinaria de petróleo. A ciência de dados e o big data evoluíram a partir da estatística e do gerenciamento tradicional de dados, mas agora são considerados disciplinas distintas.

As características do big data são frequentemente chamadas de três Vs:

- *Volume* — *Quantos* dados existem?
- *Variedade* — *Quão* diversos são os diferentes tipos de dados?
- *Velocidade* — *A* que velocidade os novos dados são gerados?

Freqüentemente, essas características são complementadas com um quarto V, a veracidade: quão precisos são os dados? Essas quatro propriedades tornam o big data diferente dos dados encontrados em ferramentas tradicionais de gerenciamento de dados. Consequentemente, os desafios que eles trazem podem ser sentida em quase todos os aspectos: captura de dados, curadoria, armazenamento, pesquisa, compartilhamento, transferência e visualização. Além disso, o big data exige técnicas especializadas para extrair os insights.

A ciência de dados é uma extensão evolutiva da estatística capaz de lidar com o enormes quantidades de dados produzidos hoje. Ele adiciona métodos da ciência da computação a o repertório de estatísticas. Numa nota de pesquisa de Laney e Kart, *Emerging Role of o Cientista de Dados e a Arte da Ciência de Dados,* os autores examinaram centenas de descrições de cargos para cientista de dados, estatístico e analista de BI (Business Intelligence) para detectar as diferenças entre esses títulos. As principais coisas que diferenciam um cientista de dados de um estatístico são a capacidade de trabalhar com big data e a experiência em aprendizado de máquina, computação e construção de algoritmos. Suas ferramentas tendem a diferir também, com as descrições de cargos de cientistas de dados mencionando com mais frequência a capacidade de use Hadoop, Pig, Spark, R, Python e Java, entre outros. Não se preocupe se você sentir intimidado por esta lista; a maioria deles será gradualmente introduzida neste livro, embora nos concentremos em Python. Python é uma ótima linguagem para ciência de dados porque tem muitas bibliotecas de ciência de dados disponíveis e é amplamente apoiada por especialistas Programas. Por exemplo, quase todos os bancos de dados NoSQL populares possuem uma interface específica para Python. API. Devido a esses recursos e à capacidade de criar protótipos rapidamente com Python enquanto mantendo um desempenho aceitável, sua influência está crescendo constantemente no mundo da ciência de dados.

À medida que a quantidade de dados continua a crescer e a necessidade de aproveitá-los torna-se mais importante, todo cientista de dados encontrará projetos de big data em todo o mundo. sua carreira.

1.1 Benefícios e utilizações da ciência de dados e big data

A ciência de dados e o big data são usados em quase todos os lugares, tanto em ambientes comerciais quanto não comerciais. O número de casos de uso é vasto e os exemplos que forneceremos ao longo deste livro apenas arranhamos a superfície das possibilidades.

Empresas comerciais em quase todos os setores usam ciência de dados e big data para obter insights sobre seus clientes, processos, equipe, conclusão e produtos. Muitos empresas usam a ciência de dados para oferecer aos clientes uma melhor experiência de usuário, bem como para venda cruzada, venda incrementada e personalize suas ofertas. Um bom exemplo disso é o Google AdSense, que coleta dados de usuários da Internet para que mensagens comerciais relevantes possam corresponder à pessoa que navega na Internet. MaxPoint (http://maxpoint.com/us)

é outro exemplo de publicidade personalizada em tempo real. Profissionais de recursos humanos usam análise de pessoas e mineração de texto para selecionar candidatos, monitorar o humor de funcionários e estudar redes informais entre colegas de trabalho. A análise de pessoas é o tema central do livro *Moneyball: The Art of Winning an Unfair Game.* No livro (e filme) vimos que o processo tradicional de observação do beisebol americano era aleatório e substituí-lo por sinais correlacionados mudou tudo. Confiando em estatísticas permitiu-lhes contratar os jogadores certos e colocá-los contra os adversários onde eles teria a maior vantagem. As instituições financeiras usam a ciência de dados para prever mercados de ações, determinar o risco de emprestar dinheiro e aprender como atrair novos clientes para os seus serviços. No momento em que este livro foi escrito, pelo menos 50% das negociações em todo o mundo eram realizadas automaticamente por máquinas baseadas em algoritmos desenvolvidos por *quants,* como costumam ser chamados os cientistas de dados que trabalham em algoritmos de negociação, com o ajuda de técnicas de big data e ciência de dados.

As organizações governamentais também estão conscientes do valor dos dados. Muitos governos as organizações não apenas contam com cientistas de dados internos para descobrir informações valiosas, mas também compartilham seus dados com o público. Você pode usar esses dados para obter insights ou construir aplicativos baseados em dados. *Data.gov* é apenas um exemplo; é a casa dos EUA Dados abertos do governo. Um cientista de dados em uma organização governamental começa a trabalhar em diversos projetos, como detecção de fraudes e outras atividades criminosas ou otimização financiamento do projeto. Um exemplo bem conhecido foi fornecido por Edward Snowden, que vazou documentos internos da Agência de Segurança Nacional Americana e da Sede de Comunicações do Governo Britânico que mostram claramente como eles usaram a ciência de dados e big data para monitorar milhões de indivíduos. Essas organizações coletaram 5 bilhões de registros de dados de aplicativos amplamente difundidos, como Google Maps, Angry Birds, e-mail e mensagens de texto, entre muitas outras fontes de dados. Em seguida, aplicaram técnicas de ciência de dados para destilar informações.

As organizações não governamentais (ONG) também conhecem bem a utilização de dados. Eles usá-lo para arrecadar dinheiro e defender suas causas. O Fundo Mundial para a Vida Selvagem (WWF), por por exemplo, emprega cientistas de dados para aumentar a eficácia de sua arrecadação de fundos esforços. Muitos cientistas de dados dedicam parte do seu tempo ajudando ONGs, porque as ONGs muitas vezes não têm recursos para coletar dados e empregar cientistas de dados. DataKind é um um grupo de cientistas de dados que dedica seu tempo ao benefício da humanidade.

As universidades usam a ciência de dados em suas pesquisas, mas também para aprimorar a experiência de estudo de seus alunos. A ascensão dos cursos on-line abertos e massivos (MOOC) produz um muitos dados, o que permite às universidades estudar como esse tipo de aprendizagem pode complementar as aulas tradicionais. Os MOOCs são um recurso inestimável se você deseja se tornar um especialista em dados cientista e profissional de big data, então dê uma olhada em alguns dos mais conhecidos: Coursera, Udacity e edX. O cenário do big data e da ciência de dados muda rapidamente, e os MOOCs permitem que você se mantenha atualizado seguindo cursos das melhores universidades. Se você ainda não os conhece, reserve um tempo para fazê-lo agora; você vai amá-los como nós temos.

1.2 Facetas dos dados

Na ciência de dados e no big data, você encontrará muitos tipos diferentes de dados, e cada um deles tende a exigir ferramentas e técnicas diferentes. As principais categorias de dados são estas:

ÿ Estruturado
ÿ Não estruturado
ÿ Linguagem natural ÿ
Gerada por máquina ÿ
Baseada em gráficos
ÿ Áudio, vídeo e imagens ÿ Streaming

Vamos explorar todos esses tipos de dados interessantes.

1.2.1 Dados estruturados

Dados estruturados são dados que dependem de um modelo de dados e residem em um campo fixo dentro de um registro. Como tal, muitas vezes é fácil armazenar dados estruturados em tabelas dentro de bases de dados ou ficheiros Excel (figura 1.1). SQL, ou Structured Query Language, é a forma preferida de gerenciar e consultar dados que residem em bancos de dados. Você também pode encontrar dados estruturados que podem dificultar seu armazenamento em um banco de dados relacional tradicional. Dados hierárquicos, como uma árvore genealógica, são um exemplo.

Porém, o mundo não é feito de dados estruturados; é imposto por humanos e máquinas. Mais frequentemente, os dados vêm desestruturados.

	Indicator ID	Dimension List	Timeframe	Numeric Value	Missing Value Flag	Confidence Inte
2	214390830	Total (Age-adjusted)	2008	74.6%		73.8%
3	214390833	Aged 18-44 years	2008	59.4%		58.0%
4	214390831	Aged 18-24 years	2008	37.4%		34.6%
5	214390832	Aged 25-44 years	2008	66.9%		65.5%
6	214390836	Aged 45-64 years	2008	88.6%		87.7%
7	214390834	Aged 45-54 years	2008	86.3%		85.1%
8	214390835	Aged 55-64 years	2008	91.5%		90.4%
9	214390840	Aged 65 years and over	2008	94.6%		93.8%
10	214390837	Aged 65-74 years	2008	93.6%		92.4%
11	214390838	Aged 75-84 years	2008	95.6%		94.4%
12	214390839	Aged 85 years and over	2008	96.0%		94.0%
13	214390841	Male (Age-adjusted)	2008	72.2%		71.1%
14	214390842	Female (Age-adjusted)	2008	76.8%		75.9%
15	214390843	White only (Age-adjusted)	2008	73.8%		72.9%
16	214390844	Black or African American only (Age-adjusted)	2008	77.0%		75.0%
17	214390845	American Indian or Alaska Native only (Age-adjusted)	2008	66.5%		57.1%
18	214390846	Asian only (Age-adjusted)	2008	80.5%		77.7%
19	214390847	Native Hawaiian or Other Pacific Islander only (Age-adjusted)	2008	DSU		
20	214390848	2 or more races (Age-adjusted)	2008	75.6%		69.6%

Figura 1.1 Uma tabela do Excel é um exemplo de dados estruturados.

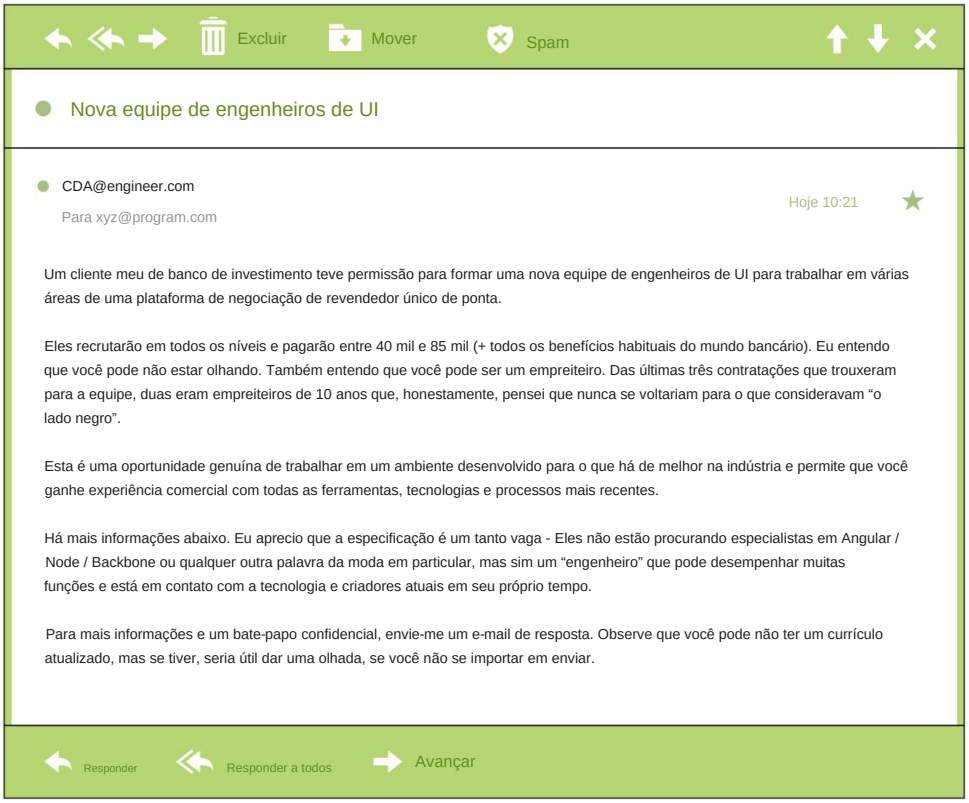


Figura 1.2 O e-mail é simultaneamente um exemplo de dados não estruturados e de dados em linguagem natural.

1.2.2 Dados não estruturados

Dados não estruturados são dados que não são fáceis de ajustar em um modelo de dados porque o conteúdo é específico do contexto ou variado. Um exemplo de dados não estruturados é o seu e-mail normal (figura 1.2). Embora o e-mail contenha elementos estruturados como remetente, título e corpo do texto, é um desafio encontrar o número de pessoas que escreveram uma reclamação por e-mail sobre um funcionário específico porque existem muitas maneiras de se referir a uma pessoa, por exemplo. Os milhares de idiomas e dialetos diferentes por aí complicam ainda mais isso.

Um e-mail escrito por humanos, conforme mostrado na figura 1.2, também é um exemplo perfeito de dados em linguagem natural.

1.2.3 Linguagem natural

A linguagem natural é um tipo especial de dados não estruturados; é um desafio de processar porque requer conhecimento de técnicas e linguística específicas de ciência de dados.

A comunidade de processamento de linguagem natural teve sucesso no reconhecimento de entidades, reconhecimento de tópicos, resumo, conclusão de texto e análise de sentimento, mas modelos treinados em um domínio não generalizam bem para outros domínios. Mesmo as técnicas mais modernas não são capazes de decifrar o significado de cada trecho de texto. Porém, isso não deveria ser uma surpresa: os humanos também lutam com a linguagem natural. É ambíguo por natureza. O próprio conceito de significado é questionável aqui. Faça com que duas pessoas ouçam a mesma conversa. Eles terão o mesmo significado? O significado das mesmas palavras pode variar quando vindas de alguém chateado ou alegre.

1.2.4 Dados gerados por máquina

Dados gerados por máquina são informações criadas automaticamente por um computador, processo, aplicativo ou outra máquina sem intervenção humana. Os dados gerados por máquinas estão a tornar-se um importante recurso de dados e continuarão a fazê-lo. A Wikibon previu que o valor de mercado da *Internet industrial* (um termo cunhado pela Frost & Sullivan para se referir à integração de maquinaria física complexa com sensores e software em rede) será de aproximadamente 540 mil milhões de dólares em 2020. IDC (International Data Corpo- ração) estimou que haverá 26 vezes mais coisas conectadas do que pessoas em 2020. Esta rede é comumente chamada de *internet das coisas.*

A análise de dados de máquinas conta com ferramentas altamente escaláveis, devido ao seu alto volume e velocidade. Exemplos de dados de máquina são logs de servidores web, registros de detalhes de chamadas, logs de eventos de rede e telemetria (figura 1.3).

```
CSIPERF:TXCOMMIT;313236
2014-11-28 11:36:13, Info                  CSI    00000153 Creating NT transaction (seq
69), objectname [6]"(null)"
2014-11-28 11:36:13, Info                  CSI    00000154 Created NT transaction (seq 69)
result 0x00000000, handle @0x4e54
2014-11-28 11:36:13, Info                  CSI    00000155@2014/11/28:10:36:13.471
Beginning NT transaction commit...
2014-11-28 11:36:13, Info                  CSI    00000156@2014/11/28:10:36:13.705 CSI perf
trace:
CSIPERF:TXCOMMIT;273983
2014-11-28 11:36:13, Info                  CSI    00000157 Creating NT transaction (seq
70), objectname [6]"(null)"
2014-11-28 11:36:13, Info                  CSI    00000158 Created NT transaction (seq 70)
result 0x00000000, handle @0x4e5c
2014-11-28 11:36:13, Info                  CSI    00000159@2014/11/28:10:36:13.764
Beginning NT transaction commit...
2014-11-28 11:36:14, Info                  CSI    0000015a@2014/11/28:10:36:14.094 CSI perf
trace:
CSIPERF:TXCOMMIT;386259
2014-11-28 11:36:14, Info                  CSI    0000015b Creating NT transaction (seq
71), objectname [6]"(null)"
2014-11-28 11:36:14, Info                  CSI    0000015c Created NT transaction (seq 71)
result 0x00000000, handle @0x4e5c
2014-11-28 11:36:14, Info                  CSI    0000015d@2014/11/28:10:36:14.106
Beginning NT transaction commit...
2014-11-28 11:36:14, Info                  CSI    0000015e@2014/11/28:10:36:14.428 CSI perf
trace:
CSIPERF:TXCOMMIT;375581
```

Figura 1.3 Exemplo de dados gerados por máquina

Os dados da máquina mostrados na figura 1.3 se encaixariam perfeitamente em um modelo clássico estruturado em tabela.
base de dados. Esta não é a melhor abordagem para dados altamente interconectados ou “em rede”,
onde as relações entre entidades têm um papel valioso a desempenhar.

1.2.5 Dados baseados em gráficos ou de rede

“Dados gráficos” pode ser um termo confuso porque qualquer dado pode ser mostrado em um gráfico.
“Gráfico”, neste caso, aponta para *a teoria matemática dos grafos.* Na teoria dos grafos, um gráfico é um
estrutura matemática para modelar relações de pares entre objetos. Gráfico ou
dados de rede são, em resumo, dados que enfocam o relacionamento ou adjacência de objetos.
As estruturas gráficas usam nós, arestas e propriedades para representar e armazenar dados gráficos.
Dados baseados em gráficos são uma forma natural de representar redes sociais, e sua estrutura
permite calcular métricas específicas, como a influência de uma pessoa e o
caminho mais curto entre duas pessoas.

Exemplos de dados baseados em gráficos podem ser encontrados em muitos sites de mídia social
(figura 1.4). Por exemplo, no LinkedIn você pode ver quem você conhece em qual empresa.
Sua lista de seguidores no Twitter é outro exemplo de dados baseados em gráficos. O poder e
a sofisticação vem de vários gráficos sobrepostos dos mesmos nós. Por exemplo, imagine as arestas
de conexão aqui para mostrar “amigos” no Facebook. Imagine
outro gráfico com as mesmas pessoas que conecta colegas de negócios via LinkedIn.
Imagine um terceiro gráfico baseado nos interesses de filmes na Netflix. Sobrepondo os três
gráficos de aparência diferente tornam possíveis perguntas mais interessantes.

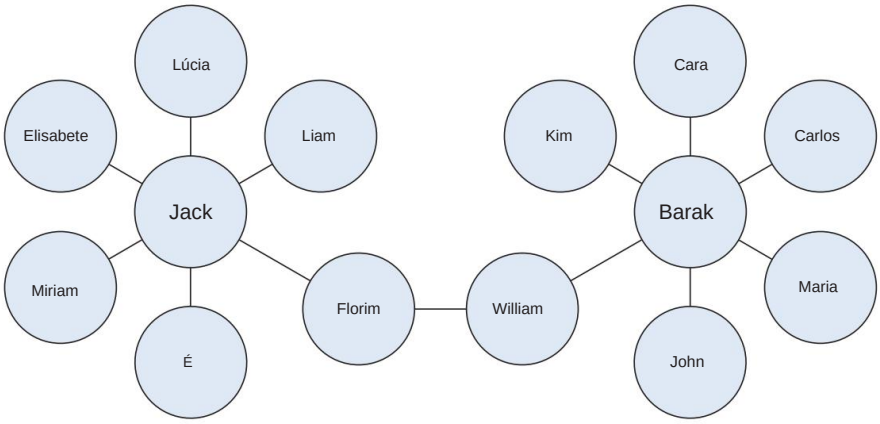


Figura 1.4 Amigos em uma rede social são um exemplo de dados baseados em gráficos.

Bancos de dados gráficos são usados para armazenar dados baseados em gráficos e são consultados com especialistas
linguagens de consulta como SPARQL.

Os dados gráficos apresentam seus desafios, mas para um aditivo e
dados de imagem, pode ser ainda mais difícil.

1.2.6 Áudio, imagem e vídeo

Áudio, imagem e vídeo são tipos de dados que apresentam desafios específicos para um cientista de dados. Tarefas que são triviais para os humanos, como reconhecer objetos em imagens, tornam-se desafiadoras para os computadores. A MLBAM (Major League Baseball Advanced Media) anunciou em 2014 que aumentaria a captura de vídeo para aproximadamente 7 TB por jogo para fins de análise ao vivo no jogo. Câmeras de alta velocidade nos estádios capturarão os movimentos da bola e dos atletas para calcular em tempo real, por exemplo, o caminho percorrido por um defensor em relação a duas linhas de base.

Recentemente, uma empresa chamada DeepMind conseguiu criar um algoritmo capaz de aprender a jogar videogame. Este algoritmo toma a tela do vídeo como entrada e aprende a interpretar tudo por meio de um processo complexo de aprendizado profundo. É um feito notável que levou o Google a comprar a empresa para seus próprios planos de desenvolvimento de Inteligência Artificial (IA) . O algoritmo de aprendizagem absorve os dados à medida que são produzidos pelo jogo de computador; é streaming de dados.

1.2.7 Transmissão de dados

Embora o streaming de dados possa assumir quase qualquer uma das formas anteriores, ele possui uma propriedade extra. Os dados fluem para o sistema quando um evento acontece, em vez de serem carregados em um armazenamento de dados em lote. Embora este não seja realmente um tipo de dado diferente, nós o tratamos aqui como tal porque você precisa adaptar seu processo para lidar com esse tipo de informação.

Exemplos são as “tendências” no Twitter, eventos esportivos ou musicais ao vivo e o mercado de ações.

1.3 O processo de ciência de dados

O processo de ciência de dados normalmente consiste em seis etapas, como você pode ver no mapa mental da Figura 1.5. Iremos apresentá-los brevemente aqui e tratá-los com mais detalhes em Capítulo 2.

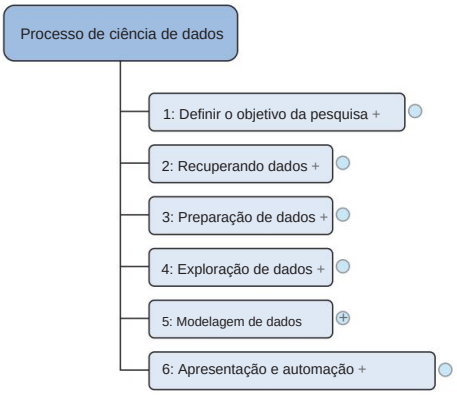


Figura 1.5 O processo de ciência de dados

1.3.1 Definição do objetivo da pesquisa

A ciência de dados é aplicada principalmente no contexto de uma organização. Quando a empresa solicitar que você execute um projeto de ciência de dados, primeiro você preparará um termo de abertura do projeto. Esta carta contém informações como o que você vai pesquisar como a empresa se beneficia com isso, quais dados e recursos você precisa, um cronograma e resultados.

Ao longo deste livro, o processo de ciência de dados será aplicado a estudos de caso maiores e você terá uma ideia dos diferentes objetivos de pesquisa possíveis.

1.3.2 Recuperando dados

A segunda etapa é coletar dados. Você declarou no termo de abertura do projeto quais dados você precisa e onde pode encontrá-los. Nesta etapa você garante que pode utilizar os dados em seu programa, o que significa verificar a existência, a qualidade e o acesso aos dados.
Os dados também podem ser entregues por empresas terceirizadas e assumem vários formatos, desde planilhas Excel até diferentes tipos de bancos de dados.

1.3.3 Preparação de dados

A coleta de dados é um processo sujeito a erros; nesta fase você aprimora a qualidade dos dados e os prepara para uso nas etapas subsequentes. Esta fase consiste em três subfases: *a limpeza de dados* remove valores falsos de uma fonte de dados e inconsistências entre fontes de dados, *a integração de dados* enriquece as fontes de dados combinando informações de múltiplas fontes de dados, e *a transformação de dados* garante que os dados estejam em uma forma adequada. formato para uso em seus modelos.

1.3.4 Exploração de dados

A exploração de dados se preocupa em construir uma compreensão mais profunda de seus dados. Você tenta entender como as variáveis interagem entre si, a distribuição dos dados e se há valores discrepantes. Para conseguir isso, você usa principalmente estatísticas descritivas, técnicas visuais e modelagem simples. Esta etapa geralmente é conhecida pela abreviatura EDA, para Exploratory Data Analysis.

1.3.5 Modelagem de dados ou construção de modelo

Nesta fase você usa modelos, conhecimento de domínio e insights sobre os dados encontrados nas etapas anteriores para responder à pergunta de pesquisa. Você seleciona uma técnica nas áreas de estatística, aprendizado de máquina, pesquisa operacional e assim por diante. A construção de um modelo é um processo iterativo que envolve a seleção das variáveis para o modelo, a execução do modelo e o diagnóstico do modelo.

1.3.6 Apresentação e automação

Por fim, você apresenta os resultados ao seu negócio. Esses resultados podem assumir diversas formas, desde apresentações até relatórios de pesquisa. Às vezes, você precisará automatizar a execução do processo porque a empresa desejará usar os insights obtidos em outro projeto ou permitir que um processo operacional use o resultado do seu modelo.

UM PROCESSO ITERATIVO A descrição anterior do processo de ciência de dados dá a impressão de que você percorre esse processo de forma linear, mas, na realidade, muitas vezes é preciso recuar e retrabalhar certas descobertas. Para Por exemplo, você pode encontrar valores discrepantes na fase de exploração de dados que apontam para erros de importação de dados. Como parte do processo de ciência de dados, você obtém ganhos incrementais insights, o que pode levar a novas questões. Para evitar retrabalho, certifique-se de que você avalia a questão comercial de forma clara e completa no início.

Agora que entendemos melhor o processo, vamos dar uma olhada nas tecnologias.

1.4 O ecossistema de big data e a ciência de dados

Atualmente existem muitas ferramentas e estruturas de big data, e é fácil se perder porque novas tecnologias aparecem rapidamente. É muito mais fácil quando você percebe que o big data O ecossistema pode ser agrupado em tecnologias que possuem objetivos e funcionalidades semelhantes, que discutiremos nesta seção. Os cientistas de dados usam muitas tecnologias diferentes, mas não todas; dedicaremos um capítulo separado aos dados mais importantes aulas de tecnologia científica. O mapa mental da figura 1.6 mostra os componentes do ecossistema de big data e onde pertencem as diferentes tecnologias.

Vejamos os diferentes grupos de ferramentas neste diagrama e vejamos o que cada um faz. Começaremos com sistemas de arquivos distribuídos.

1.4.1 Sistemas de arquivos distribuídos

Um *sistema de arquivos distribuído* é semelhante a um sistema de arquivos normal, exceto pelo fato de ser executado em vários servidores de uma só vez. Por ser um sistema de arquivos, você pode fazer quase todas as mesmas coisas que faria faça em um sistema de arquivos normal. Ações como armazenar, ler e excluir arquivos e adicionar segurança aos arquivos está no centro de todo sistema de arquivos, incluindo o distribuído um. Os sistemas de arquivos distribuídos têm vantagens significativas:

- ÿ Eles podem armazenar arquivos maiores que qualquer disco de computador.
- ÿ Os arquivos são replicados automaticamente em vários servidores para redundância ou operações paralelas, ocultando ao usuário a complexidade de fazer isso.
- ÿ O sistema é facilmente dimensionado: você não está mais limitado pela memória ou armazenamento restrições de um único servidor.

No passado, a escala era aumentada ao mover tudo para um servidor com mais memória, armazenamento e uma CPU melhor (escala vertical). Hoje em dia você pode adicionar outro pequeno servidor (escala horizontal). Este princípio torna o potencial de escala virtualmente ilimitado.

O sistema de arquivos distribuído mais conhecido atualmente é o *Hadoop File System (HDFS).* É uma implementação de código aberto do Google File System. Neste livro nos concentramos no sistema de arquivos Hadoop porque é o mais comum em uso. Entretanto, existem muitos outros sistemas de arquivos distribuídos: *Red Hat Cluster File System, Ceph File System* e *Tachyon File System,* para citar apenas três.

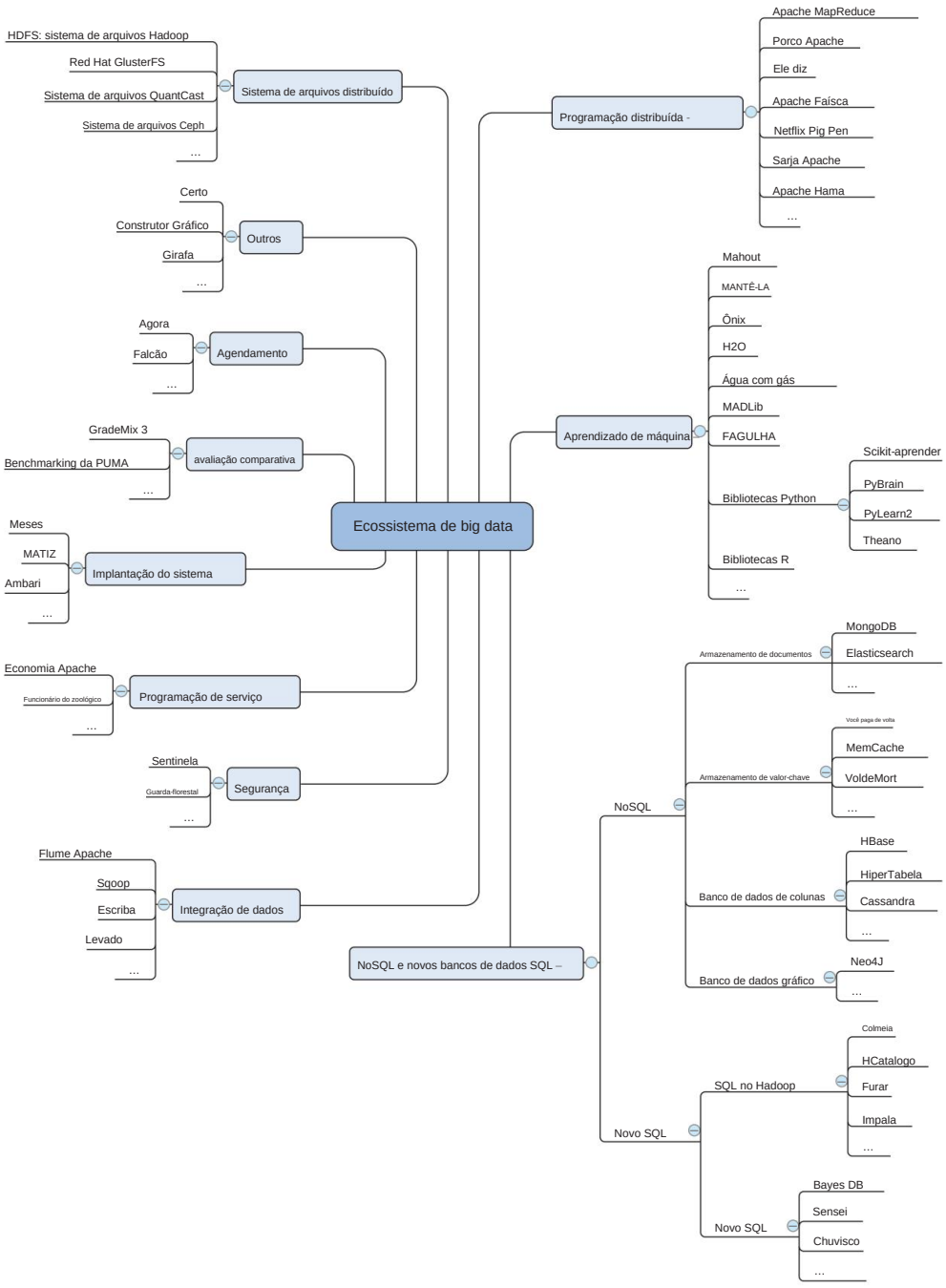


Figura 1.6 As tecnologias de big data podem ser classificadas em alguns componentes principais.

1.4.2 Estrutura de programação distribuída

Depois de armazenar os dados no sistema de arquivos distribuído, você deseja explorá-los. Um aspecto importante de trabalhar em um disco rígido distribuído é que você não moverá seus dados para o programa, mas sim moverá seu programa para os dados. Ao começar do zero com uma linguagem de programação normal de uso geral, como C, Python ou Java, você precisa lidar com as complexidades que acompanham a programação distribuída, como reiniciar tarefas que falharam, rastrear os resultados dos diferentes subprocessos , e assim por diante. Felizmente, a comunidade de código aberto desenvolveu muitas estruturas para lidar com isso para você, e elas proporcionam uma experiência muito melhor ao trabalhar com dados distribuídos e lidar com muitos dos desafios que eles acarretam.

1.4.3 Estrutura de integração de dados

Depois de instalar um sistema de arquivos distribuído, você precisará adicionar dados. Você precisa mover dados de uma fonte para outra, e é aqui que funcionam as estruturas de integração de dados, como Apache Sqoop e Apache Flume Excel. O processo é semelhante a um processo de extração, transformação e carregamento em um data warehouse tradicional.

1.4.4 Estruturas de aprendizado de máquina

Quando você tiver os dados prontos, é hora de extrair os cobiçados insights. É aqui que você confia nas áreas de aprendizado de máquina, estatística e matemática aplicada. Antes da Segunda Guerra Mundial, tudo precisava ser calculado manualmente, o que limitava severamente as possibilidades de análise de dados. Após a Segunda Guerra Mundial, foram desenvolvidos computadores e computação científica. Um único computador poderia fazer todas as contagens e cálculos e um mundo de oportunidades se abriria. Desde essa descoberta, as pessoas só precisam derivar as fórmulas matemáticas, escrevê-las em um algoritmo e carregar seus dados. Com a enorme quantidade de dados disponíveis hoje em dia, um computador não consegue mais lidar sozinho com a carga de trabalho. Na verdade, vários algoritmos desenvolvidos no milénio anterior nunca terminariam antes do fim do universo, mesmo que pudéssemos usar todos os computadores disponíveis na Terra. Isso tem a ver com complexidade de tempo (https://en.wikipedia.org/wiki/Time_complexity). Um exemplo é tentar quebrar uma senha testando todas as combinações possíveis. Um exemplo pode ser encontrado em http://stackoverflow.com/questions/7055652/real-world-example-of-exponencial-time-complexity . Um dos maiores problemas com os algoritmos antigos é que eles não são bem dimensionados. Com a quantidade de dados que precisamos analisar hoje, isso se torna problemático, e são necessárias estruturas e bibliotecas especializadas para lidar com essa quantidade de dados. A biblioteca de aprendizado de máquina mais popular para Python é a Scikit-learn. É uma ótima caixa de ferramentas de aprendizado de máquina e a usaremos mais adiante neste livro. Existem, é claro, outras bibliotecas Python:

- *PyBrain para redes neurais* — Redes neurais são algoritmos de aprendizagem que imitam o cérebro humano no aprendizado de mecânica e complexidade. As redes neurais são frequentemente consideradas avançadas e caixa preta.

ÿ *NLTK ou Natural Language Toolkit—Como* o nome sugere, seu foco é trabalhar com linguagem natural. É uma extensa biblioteca que vem com vários corpus de texto para ajudá-lo a modelar seus próprios dados.

ÿ Pylearn2 – Outra caixa de ferramentas de aprendizado de máquina, mas um pouco menos madura que o Scikit-learn.

ÿ *TensorFlow – uma* biblioteca Python para aprendizado profundo fornecida pelo Google.

O cenário não termina com as bibliotecas Python, é claro. Spark é um novo mecanismo de aprendizado de máquina licenciado pelo Apache, especializado em aprendizado de máquina em tempo real. Isso é vale a pena dar uma olhada e você pode ler mais sobre isso em http://spark.apache.org/.

1.4.5 Bancos de dados NoSQL

Se precisar armazenar grandes quantidades de dados, você precisará de um software especializado em gerenciar e consultar esses dados. Tradicionalmente, esse tem sido o campo de atuação dos bancos de dados relacionais, como Oracle SQL, MySQL, Sybase IQ e outros. Enquanto eles ainda estão a tecnologia ideal para muitos casos de uso, novos tipos de bancos de dados surgiram sob o agrupamento de bancos de dados NoSQL .

O nome deste grupo pode ser enganoso, pois “Não” neste contexto significa “Não Apenas." A falta de funcionalidade em SQL não é o maior motivo para a mudança de paradigma, e muitos dos bancos de dados NoSQL implementaram eles próprios uma versão do SQL . Mas os bancos de dados tradicionais tinham deficiências que não lhes permitiam uma boa escalabilidade. Ao resolver vários dos problemas dos bancos de dados tradicionais, os bancos de dados NoSQL permitem uma virtualmente crescimento infinito de dados. Estas deficiências dizem respeito a todas as propriedades do big data: a sua o poder de armazenamento ou processamento não pode escalar além de um único nó e eles não têm como lidar com streaming, gráfico ou formas não estruturadas de dados.

Muitos tipos diferentes de bancos de dados surgiram, mas eles podem ser categorizados em seguintes tipos:

ÿ *Bancos de dados de colunas — os dados* são armazenados em colunas, o que permite que os algoritmos realizem consultas muito mais rápidas. As tecnologias mais recentes usam armazenamento em células. Semelhante a uma mesa estruturas ainda são importantes.

ÿ Armazenamentos *de documentos* – Os armazenamentos de documentos não usam mais tabelas, mas armazenam todas as observações em um documento. Isso permite um esquema de dados muito mais flexível.

ÿ *Streaming* de dados — os dados são coletados, transformados e agregados não em lotes mas em tempo real. Embora o tenhamos categorizado aqui como um banco de dados para ajudá-lo seleção de ferramentas, é mais um tipo específico de problema que impulsionou a criação de tecnologias como Storm.

ÿ *Armazenamentos de valores-chave — os dados* não são armazenados em uma tabela; em vez disso, você atribui uma chave para cada valor, como org.marketing.sales.2015: 20000. Isso é bem dimensionado, mas coloca quase toda a implementação no desenvolvedor.

ÿ *SQL no* Hadoop — as consultas em lote no Hadoop estão em uma linguagem semelhante a SQL que usa a estrutura de redução de mapa em segundo plano.

ÿ *Novo* SQL — Esta classe combina a escalabilidade dos bancos de dados NoSQL com o vantagens dos bancos de dados relacionais. Todos eles possuem uma interface SQL e um modelo de dados relacional.

ÿ *Bancos de dados gráficos – Nem* todo problema é melhor armazenado em uma tabela. Problemas específicos são traduzidos mais naturalmente na teoria dos grafos e armazenados em bancos de dados de grafos. Um exemplo clássico disso é uma rede social.

1.4.6 Ferramentas de agendamento

As ferramentas de agendamento ajudam a automatizar tarefas repetitivas e acionar trabalhos com base em eventos, como adicionar um novo arquivo a uma pasta. Eles são semelhantes a ferramentas como CRON no Linux, mas são desenvolvidos especificamente para big data. Você pode usá-los, por exemplo, para iniciar uma tarefa MapReduce sempre que um novo conjunto de dados estiver disponível em um diretório.

1.4.7 Ferramentas de benchmarking

Esta classe de ferramentas foi desenvolvida para otimizar sua instalação de big data, fornecendo conjuntos de criação de perfil padronizados. Um conjunto de criação de perfil é obtido de um conjunto representativo de trabalhos de big data. O benchmarking e a otimização da infraestrutura e configuração de big data nem sempre são tarefas para os próprios cientistas de dados, mas para um profissional especializado na configuração da infraestrutura de TI ; portanto, eles não são abordados neste livro. Usar uma infraestrutura otimizada pode fazer uma grande diferença de custos. Por exemplo, se você conseguir ganhar 10% em um cluster de 100 servidores, você economizará no custo de 10 servidores.

1.4.8 Implantação do sistema

Configurar uma infraestrutura de big data não é uma tarefa fácil e auxiliar os engenheiros na implantação de novos aplicativos no cluster de big data é onde as ferramentas de implantação do sistema brilham. Eles automatizam amplamente a instalação e configuração de componentes de big data. Esta não é uma tarefa central de um cientista de dados.

1.4.9 Programação de serviço Suponha que

você tenha criado um aplicativo de previsão de futebol de classe mundial no Hadoop e queira permitir que outras pessoas usem as previsões feitas por seu aplicativo. No entanto, você não tem ideia da arquitetura ou tecnologia de todos que desejam usar suas previsões. As ferramentas de serviço se destacam aqui, expondo aplicações de big data a outras aplicações como um serviço. Às vezes, os cientistas de dados precisam expor seus modelos por meio de serviços. O exemplo mais conhecido é o serviço REST ; REST significa transferência de estado representacional. Muitas vezes é usado para alimentar sites com dados.

1.4.10 Segurança

Você deseja que todos tenham acesso a todos os seus dados? Você provavelmente precisará ter um controle refinado sobre o acesso aos dados, mas não deseja gerenciar isso aplicativo por aplicativo. As ferramentas de segurança de big data permitem que você tenha controle central e refinado sobre o acesso aos dados. A segurança de big data tornou-se um tema por si só, e os cientistas de dados geralmente só são confrontados com ela como consumidores de dados; raramente eles próprios implementarão a segurança. Neste livro não descrevemos como configurar a segurança em big data porque esse é um trabalho para o especialista em segurança.

1.5 Um exemplo introdutório de trabalho do Hadoop

Terminaremos este capítulo com uma pequena aplicação em um contexto de big data. Para isso usaremos uma imagem do Hortonworks Sandbox. Esta é uma máquina virtual criada pela Hortonworks para experimente alguns aplicativos de big data em uma máquina local. Mais adiante neste livro você verá como Juju facilita a instalação do Hadoop em várias máquinas.

Usaremos um pequeno conjunto de dados de salários de cargos para executar nossa primeira amostra, mas consultando um um grande conjunto de dados de bilhões de linhas seria igualmente fácil. A linguagem de consulta parecerá como SQL, mas nos bastidores um trabalho MapReduce será executado e produzirá uma tabela direta de resultados, que pode então ser transformada em um gráfico de barras. O resultado final disso o exercício se parece com a figura 1.7.

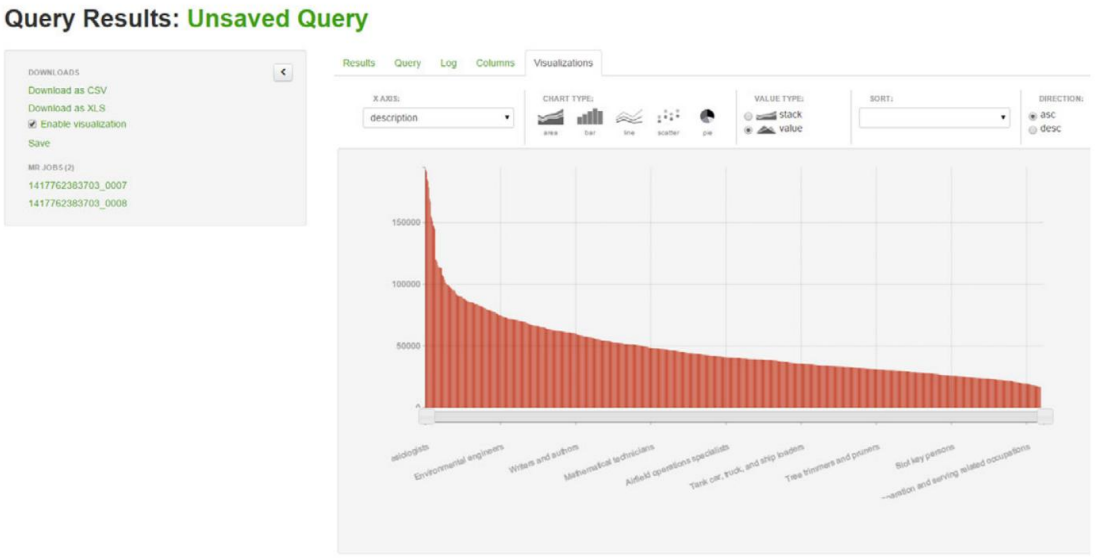


Figura 1.7 O resultado final: o salário médio por descrição de cargo

Para começar a funcionar o mais rápido possível, usamos um Hortonworks Sandbox dentro do Virtual-Box. VirtualBox é uma ferramenta de virtualização que permite executar outro sistema operacional dentro do seu próprio sistema operacional. Neste caso você pode executar o CentOS com um sistema existente Instalação do Hadoop dentro do sistema operacional instalado.

Algumas etapas são necessárias para colocar o sandbox em funcionamento no VirtualBox. Cuidado, as seguintes etapas eram aplicáveis no momento em que este capítulo foi escrito (fevereiro de 2015):

1 Baixe a imagem virtual em http://hortonworks.com/products/hortonworks-sandbox/#install .

2 Inicie o host da sua máquina virtual. O VirtualBox pode ser baixado em
https://www.virtualbox.org/wiki/Downloads.

3 Pressione CTRL+I e selecione a imagem virtual do Hortonworks.

4 Clique em Avançar.

5 Clique em Importar; depois de um tempo sua imagem deverá ser importada.

6 Agora selecione sua máquina virtual e clique em Executar.

7 Aguarde um pouco para iniciar a distribuição do CentOS com a instalação do Hadoop em execução, conforme mostrado na figura 1.8. Observe que a versão do Sandbox aqui é 2.1. Com outras versões as coisas poderiam ser um pouco diferentes.

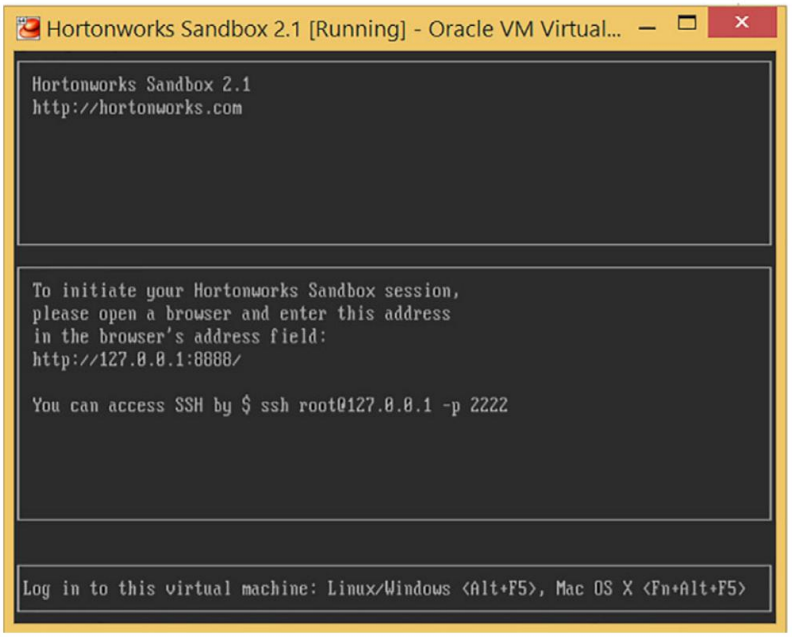


Figura 1.8 Hortonworks Sandbox rodando no VirtualBox

Você pode fazer logon diretamente na máquina ou usar SSH para fazer logon. Para este aplicativo você usará a interface da web. Aponte seu navegador para o endereço http://127.0.0.1:8000 e você será recebido com a tela mostrada na figura 1.9.

Hortonworks carregou dois conjuntos de amostras, que você pode ver no HCatalog. Basta clicar no botão HCat na tela e você verá as tabelas disponíveis (figura 1.10).

Figura 1.9 A caixa de areia Hortonworks tela de boas vindas disponível em http://127.0.0.1:8000

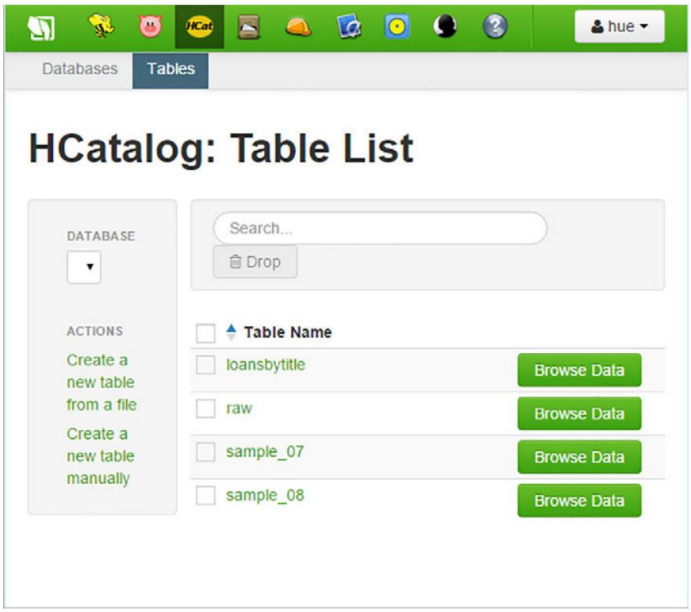
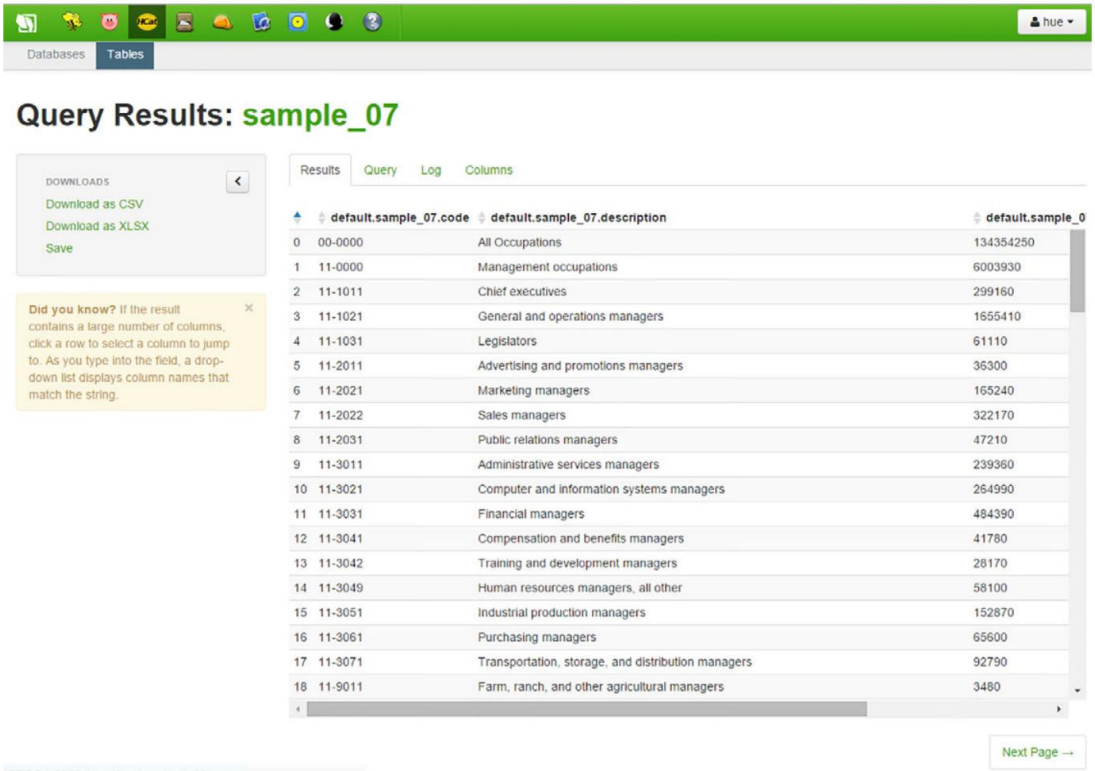


Figura 1.10 Uma lista de tabelas disponíveis no HCatalog



	default.sample_07.code	default.sample_07.description	default.sample_0
0	00-0000	All Occupations	134354250
1	11-0000	Management occupations	6003930
2	11-1011	Chief executives	299160
3	11-1021	General and operations managers	1655410
4	11-1031	Legislators	61110
5	11-2011	Advertising and promotions managers	36300
6	11-2021	Marketing managers	165240
7	11-2022	Sales managers	322170
8	11-2031	Public relations managers	47210
9	11-3011	Administrative services managers	239360
10	11-3021	Computer and information systems managers	264990
11	11-3031	Financial managers	484390
12	11-3041	Compensation and benefits managers	41780
13	11-3042	Training and development managers	28170
14	11-3049	Human resources managers, all other	58100
15	11-3051	Industrial production managers	152870
16	11-3061	Purchasing managers	65600
17	11-3071	Transportation, storage, and distribution managers	92790
18	11-9011	Farm, ranch, and other agricultural managers	3480

Figura 1.11 O conteúdo da tabela

Para ver o conteúdo dos dados, clique no botão Procurar dados próximo ao sample_07 entrada para obter a próxima tela (figura 1.11).

Parece uma mesa comum e o Hive é uma ferramenta que permite abordá-la como um banco de dados comum com SQL. Isso mesmo: no Hive você obtém seus resultados usando o HiveQL, um dialeto do SQL antigo . Para abrir o editor Beeswax HiveQL , clique no botão Beeswax no menu (figura 1.12).

Para obter seus resultados, execute a seguinte consulta:

```
Selecione descrição, avg(salary) como salário_médio do grupo sample_07 por ordem de descrição por
salário_médio desc.
```

Clique no botão Executar. O Hive traduz seu HiveQL em um job MapReduce e o executa em seu ambiente Hadoop, como você pode ver na figura 1.13.

No entanto, é melhor evitar a leitura da janela de log por enquanto. Neste ponto, é enganoso. Se esta for sua primeira consulta, poderá levar 30 segundos. O Hadoop é famoso por seu períodos de aquecimento. Essa discussão fica para mais tarde, no entanto.

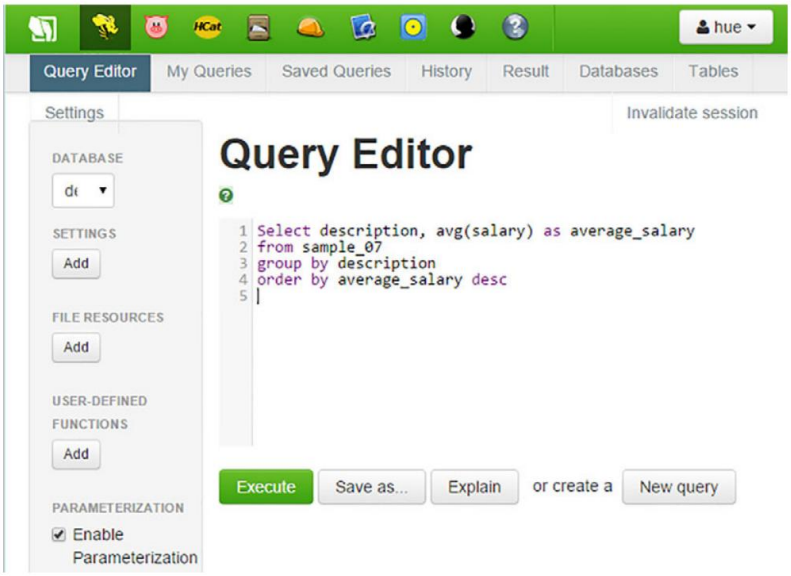


Figura 1.12 Você pode executar um comando HiveQL no editor Beeswax HiveQL. Nos bastidores, isso é traduzido em um trabalho MapReduce.

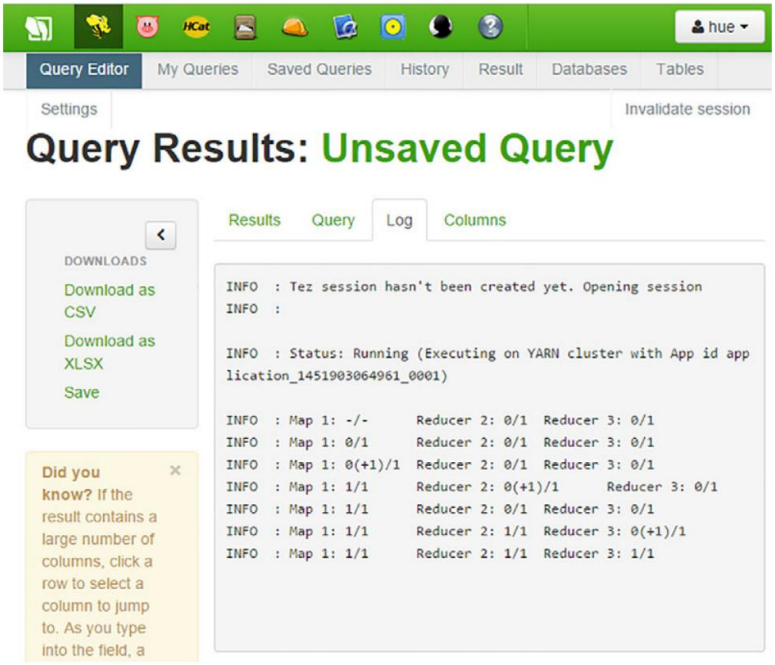


Figura 1.13 O registro mostra que seu HiveQL é traduzido em um trabalho MapReduce. Nota: Este log é da versão de fevereiro de 2015 do HDP, portanto a versão atual pode parecer um pouco diferente.

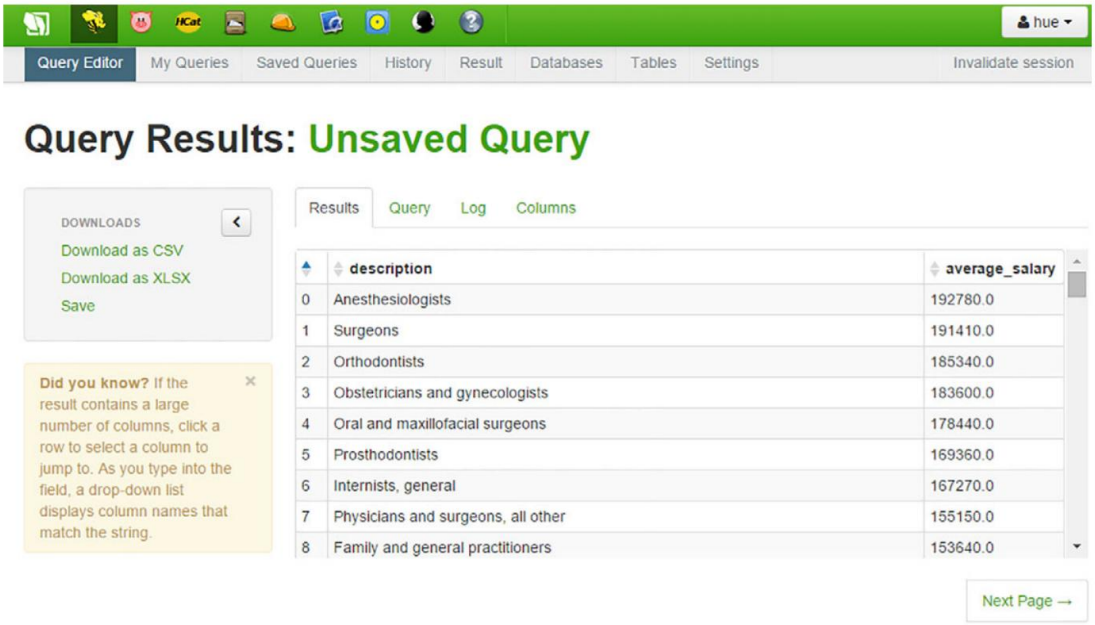


Figura 1.14 O resultado final: uma visão geral do salário médio por profissão

Depois de um tempo o resultado aparece. Ótimo trabalho! A conclusão disso, conforme mostra a figura 1.14, é que cursar medicina é um bom investimento. Surpreso?

Com esta tabela concluímos nosso tutorial introdutório ao Hadoop.

Embora este capítulo tenha sido apenas o começo, às vezes pode ter parecido um pouco opressor. Recomenda-se deixar como está por enquanto e voltar aqui novamente quando todos os conceitos estiverem completamente explicados. A ciência de dados é um campo amplo, por isso vem com um vocabulário amplo. Esperamos dar a você um vislumbre da maior parte disso durante nosso tempo juntos. Depois, você escolhe e aprimora suas habilidades na direção que mais lhe interessa. É disso que se trata “Introduzindo a Ciência de Dados” e esperamos que você aproveite a viagem conosco.

1.6 Resumo

Neste capítulo você aprendeu o seguinte:

- *Big data* é um termo geral para qualquer coleção de conjuntos de dados tão grandes ou complexos que se torna difícil processá-los usando técnicas tradicionais de gerenciamento de dados. Eles são caracterizados pelos quatro Vs: velocidade, variedade, volume e veracidade.

- *A ciência de dados* envolve o uso de métodos para analisar pequenos conjuntos de dados até os gigantescos de que trata o big data.

Resumo

ÿ Embora o *processo de ciência de dados* não seja linear, ele pode ser dividido em etapas:

1 Definir o objetivo da pesquisa

2 Coletando dados

3 Preparação de dados

4 Exploração de dados

5 Modelagem

6 Apresentação e automação

ÿ O cenário do big data é mais do que apenas o Hadoop. Consiste em muitas tecnologias diferentes que podem ser categorizadas da seguinte forma: – Sistema de arquivos – Estruturas de programação distribuída – Integração de dados – Bancos de dados

– Aprendizado de máquina

– Segurança

– Programação –

Benchmarking –

Implantação de sistema –

Programação de serviços ÿ Nem todas as categorias de big data são amplamente utilizadas por cientistas de dados. Eles se concentram principalmente no sistema de arquivos, nas estruturas de programação distribuída, nos bancos de dados e no aprendizado de máquina. Eles entram em contato com os outros componentes, mas estes são domínios de outras profissões.

ÿ Os dados podem vir em diferentes formas. As principais formas são

– Dados estruturados

– Dados não estruturados

– Dados de linguagem natural

– Dados da máquina

– Dados baseados em gráficos

– Transmissão de dados