LEARNING MADE EASY

Snowflake Special Edition

Cloud Data Science

for dummies®
A Wiley Brand

Learn all about cloud data science

Scale machine learning with near-unlimited data

Democratize data science across your organization

Brought to you by

snowflake®

David Baum

About Snowflake

Snowflake delivers the Data Cloud — a global network where thousands of organizations mobilize data with near-unlimited scale, concurrency, and performance. Inside the Data Cloud, organizations unite their siloed data, easily discover and securely share governed data, and execute diverse analytic workloads. Wherever data or users live, Snowflake delivers a single and seamless experience across multiple public clouds. Snowflake's platform is the engine that powers and provides access to the Data Cloud, creating a solution for data warehousing, data lakes, data engineering, data science, data application development, and data sharing. Join Snowflake customers, partners, and data providers already taking their businesses to new frontiers in the Data Cloud. **snowflake.com.**

Cloud Data Science

Snowflake Special Edition

by David Baum

for dummies®
A Wiley Brand

Cloud Data Science For Dummies®, Snowflake Special Edition

Published by
John Wiley & Sons, Inc.
111 River St.
Hoboken, NJ 07030-5774
www.wiley.com

ISBN 978-1-119-82939-3 (pbk); ISBN 978-1-119-82940-9 (ebk)

Manufactured in the United States of America

10 9 8 7 6 5 4 3 2 1

Publisher's Acknowledgments

Some of the people who helped bring this book to market include the following:

Development Editor: Brian Walls
Project Manager: Jen Bingham
Acquisitions Editor: Ashley Coffey
Editorial Manager: Rev Mengle
Business Development Representative: Molly Daugherty
Content Refinement specialist: Tamilmani Varadharaj
Snowflake Contributors Team: Julian Forero, Vincent Morello, Ahmad Khan, Jena Donlin, Shiyi Gu, Kristen Werner, Simon Field, Mike Klaczynski, Leslie Steere

Table of Contents

Introduction

Data science relies on scientific methods, processes, and algorithms to extract knowledge and derive insights from data. Beyond business intelligence (BI), which describes what *has* happened, data science is *predictive* and *prescriptive:* Data scientists create intelligent information systems that can predict future outcomes and automate common tasks, from fraud-detection systems that protect our bank accounts to intelligent firewalls that distinguish between benign and malicious network activity.

A growing number of organizations use data science to create machine learning (ML) models that can assess real-time conditions and make split-second decisions, such as presenting custom offers to Internet shoppers based on their past purchase histories, moment-to-moment browsing activity, and regional sales trends. Data science has broadly penetrated nearly every industry as ML technology finds its way into automobiles, factories, power grids, hospitals, distribution centers, and business processes of all types.

Here's the caveat: ML models, and hence the decisions made from those models, are only as good as the data that supports them. The more data these models ingest and the more situations they encounter, the smarter and more accurate they become. And yet managing data remains one of the field's most onerous tasks. According to an Anaconda report titled "The State of Data Science 2020: Moving from Hype Toward Maturity," data scientists spend almost two-thirds of their time collecting, preparing, and visualizing data — leaving little time to build the predictive models that put that data to work. Challenges include:

- Unifying your data from multiple silos and diverse formats to establish a single repository that multiple workgroups can easily and securely access
- Accessing near-unlimited data storage and compute power to scale your data science apps from test to production across teams and organizations
- Centralizing data governance in order to democratize your modern data science initiatives across your organization

This book provides an overview of data science and emphasizes the data foundation that organizations need in order to succeed with their data science initiatives. Read on to learn about this rapidly evolving, data-intensive field responsible for creating today's most intelligent and data-driven decision-making solutions.

Icons Used in This Book

Throughout this book, the following icons highlight tips, important points to remember, real-life examples, and more.


TIP

Guidance on how to take advantage of cloud data science in your organization.


REMEMBER

Concepts worth remembering as you immerse yourself in understanding data science.


CASE STUDY

Case studies about organizations using cloud data science to understand markets and customers, assess risks, anticipate customer needs, and more.


TECHNICAL STUFF

The jargon beneath the jargon, explained.

Beyond the Book

If you like what you read in this book, visit `www.snowflake.com` for a free trial of the Data Cloud; to obtain details about plans and pricing, view webinars, or access detailed documentation; or to get in touch with a member of the Snowflake team.

IN THIS CHAPTER

- Defining data science terms and concepts
- Moving from descriptive to predictive analytics
- Understanding the potential of data science in business and industry
- Distinguishing between narrow and general AI
- Using a cloud data platform as a foundation for data science

Chapter 1
Getting Up to Speed on Cloud Data Science

In 2006, a marketing commentator at the Association of National Advertisers named Michael Palmer made the now-famous assertion that raw data is like crude oil: It must be refined to release its potential. Just as crude oil must be converted into gas, plastic, chemicals, and other commodities, he reasoned, data must be captured, transformed, stored, processed, and analyzed to yield useful insights.

A May 2017 article in *The Economist* echoed this assertion when it pointed out that today's most valuable companies are no longer the Industrial Era's oil companies but the Digital Age's tech titans: Alphabet, Amazon, Apple, Facebook, and Microsoft. Each of these technology leaders has forged a unique path in the digital frontier, but they all have one thing in common: astute use of data, driven by advanced data science practices.

Defining Terms and Concepts

Data science depends on a broad range of software tools, algorithms, and machine learning principles to uncover business insights hidden in vast volumes of data. *Data scientists* make data valuable by collecting, transforming, and turning data into predictive and prescriptive insights.

Under the umbrella of data science, *artificial intelligence (AI)* concerns the science and engineering of creating intelligent machines, especially intelligent computer programs that can process information and produce outcomes autonomously. An AI system's goal is to solve complex problems in ways that are similar to human logic and reasoning.

REMEMBER

Machine learning (ML), a subset of AI, refers to the algorithms that are part of the *data science workflow* — specialized software programs designed to detect patterns, make correlations, identify anomalies in the data, and ultimately predict outcomes. These programs also enable computers to learn as they're exposed to new data and situations, increasing prediction accuracy as they're provided additional relevant information.

Most people can recognize the relationships among only five to ten different attributes. An ML algorithm can process thousands of attributes and features to surface unique combinations and correlations in the data. ML is a useful form of AI for business because it allows computers to apply knowledge and extract value from large data sets. For example, Netflix tracks thousands of data points among hundreds of millions of subscribers. Its algorithms can predict how people will respond to certain types of programs based on their ratings, interests, previous viewing habits, and clickstream histories.

REMEMBER

Deep learning (DL), a subset of ML, uses ML techniques to simulate human decision-making by tapping into *neural networks* that mimic a human brain's neurons and nodes. As data sets grow in richness and scope, some data science applications evolve from using ML to using DL, as illustrated in Figure 1-1.

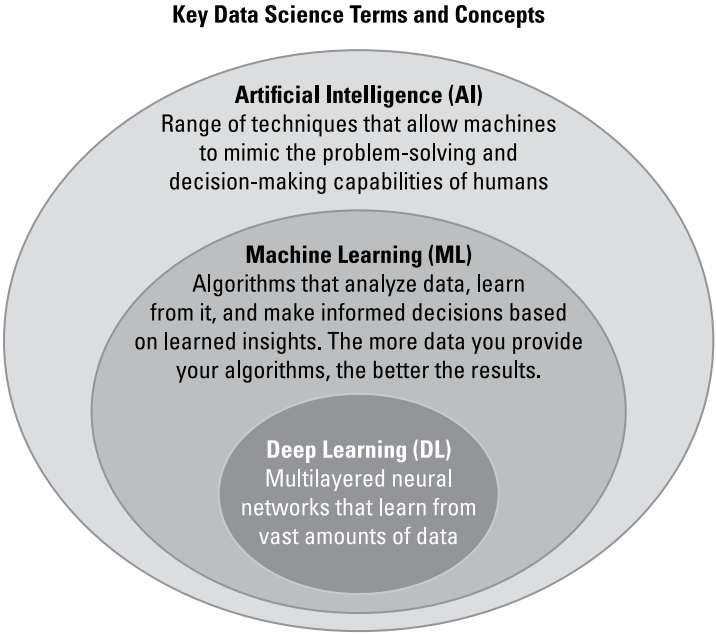

FIGURE 1-1: Data science can predict and prescribe the future.

Moving from Historical to Predictive Analytics

Traditional business intelligence (BI) tools mainly produce *descriptive analytics,* which is used for historical reporting. For example, a wholesaler might depend on a warehouse management dashboard to monitor incoming and outgoing shipments to warehouses and distribution centers. The dashboard can display metrics about stock levels, shelf space, and distribution plans, allowing users to sort and filter the data by region, category, stock keeping unit (SKU), and other variables. Warehouse managers and supply chain managers can visualize the results in charts, graphs, and reports. They can also query the data to diagnose problems, such as determining what caused the shortage in raw materials that had an impact on production timelines.

On the flip side, data science helps automate everyday tasks, from pharmaceutical applications that predict the likelihood of adverse drug interactions to public health systems that analyze the impact vaccinations have on infections. These apps enable better

decision-making by enabling the shift from descriptive to predictive analytics — from viewing a dashboard to determine *what* happened to consulting a predictive model to figure out *what will happen* and *how* the outcome is likely to evolve (see Figure 1-2).

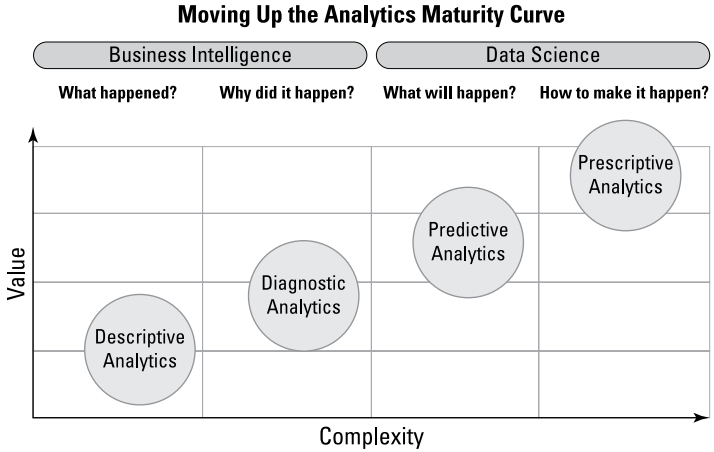

FIGURE 1-2: Data science enables the progression from descriptive and diagnostic BI to predictive and prescriptive analytics.

REMEMBER

Descriptive and *diagnostic* systems make determinations about what happened last week or last month by examining historical data. *Predictive analytics* systems assess future outcomes and conditions. They examine patterns in your data in conjunction with third-party data sources to assess probable outcomes.

Once data scientists train a model with the chosen algorithm, these systems can reach their own conclusions. In warehouse management, for example, an intelligent supply chain can track the flow of goods and materials relative to customer demand. It can simulate dozens of potential market scenarios to assess their impact on revenue and margins. By sharing data and insights, supply chain participants, manufacturers, wholesalers, retailers, and logistics service providers can collectively resolve disruptions and address inventory shortfalls before they impact customer service.

Predictive analytics and ML go hand-in-hand because predictive models typically include an ML algorithm. These systems may pull in third-party data about news, events, weather, and other factors to gauge their impact on production, planning, and replenishment scenarios. For example, Uber's ML models can predict spikes in rider demand based on regional weather conditions; the

timing of events, concerts, and conferences; and historical data about rider trends. Its models ensure drivers are directed to the right locations and automatically adjust the number of drivers on the road to meet demand and drive more revenue.

Using AI in Business and Industry

Data science isn't just the province of large, data-driven companies such as Facebook and Uber. Businesses in every industry use data to increase revenue, reduce costs, and minimize risks — and many of them have either nascent or advanced data science initiatives. Whether making product recommendations, discovering target audiences, or fine-tuning operating conditions, these ML algorithms look for patterns by comparing current and historical data from a wide variety of sources. According to a November 2020 McKinsey survey titled "The state of AI in 2020," a growing contingent of companies use AI as a tool for generating value in the form of increased revenues. In fact, half of the respondents to McKinsey's survey said their organizations had adopted AI in at least one area of their businesses.

Distinguishing Between Narrow and General AI

In the early stages of AI, hype exceeded reality by a wide margin as computer scientists set out to create general-purpose machines that could think and act like humans. The industry reached a low point in the early 1990s due to unrealistic expectations from AI researchers and the general public's skepticism. One of the primary reasons for the ensuing "AI Winter" was that AI was equated with general intelligence.

A turning point came in 1996 when IBM's Deep Blue defeated chess grandmaster Garry Kasparov. This surprise upset was a wake-up call for society and a turning point for the AI industry. The key to Deep Blue's success was not its ability to think or act like a human but its ability to master one highly constrained task: the game of chess. Similarly, in 2016, DeepMind's AlphaGo computer program beat Lee Sedol, one of the world's best Go players, in a five-game match of the board game Go.

Today, these so-called *narrow AI* systems successfully handle a huge variety of domain-specific tasks, from helping banks identify fraudulent financial transactions to helping motorists find the quickest, safest routes. Common examples of AI in industry include the following:

- **Semiconductor firms** use AI to improve the yield on the integrated circuits that power everything from communications satellites to mobile phones. In these capital-intensive fabrication centers, increasing efficiency by a fraction of a percentage point can translate to millions of dollars of savings.
- **Cybersecurity** teams use AI to translate trillions of raw data points into real-time security insights. Algorithms rank and score network-event data to flag anomalies that may indicate suspicious behavior, maximizing the efficiency of security teams and reducing IT overhead.
- **Manufacturers** use data science to collect Internet of Things (IoT) data from connected devices. ML algorithms study environmental and mechanical conditions to predict when a machine on an assembly line or shop floor requires maintenance or is about to fail.
- **Retailers** use AI to monitor supply chain interactions, predict consumer demand, and automatically replenish goods in warehouses and distribution centers, and on retail shelves.

In these and countless other business examples, access to vast amounts of relevant data is what makes ML models function. Just as AlphaGo gained skill by playing and studying thousands of Go matches, AI applications get smarter through repeated training from relevant data sets.

Enhancing Business Processes with AI

In many situations, AI-powered software programs don't replace humans: They augment their activities. Rather than fearing AI automation, forward-looking businesses see its potential to liberate workers from mundane tasks. For example, within the field of radiology, an AI program can use advanced image-recognition technology to detect worrisome features in X-rays, MRIs, and other medical scans. The program uses knowledge learned from studying millions of similar images to isolate potential

problems and anomalies. This analysis frees time for physicians to concentrate on the outliers. Rather than manually viewing a hundred chest X-rays, a radiologist can focus on the ones the computer program has singled out for further scrutiny, such as those with the presence of white spots that signify pneumonia.

As machines take on routine and repetitive tasks, people can tackle cases that require human discernment. Machines do the prework; specialists study the exceptions. For example, a growing number of call centers successfully use AI-driven chatbots to answer basic questions, allowing human support personnel to concentrate on complex or sensitive tickets.

In addition to automating many existing tasks, data science has created new positions for skilled technology workers, including data analysts, statisticians, data engineers, and software developers (see Chapter 3).

REMEMBER

ML models can learn to accomplish a huge variety of tasks. However, the learning comes from historical data and outcomes provided by a data scientist. To effectively teach models, data scientists need to clearly understand the business problem, form a hypothesis, source and transform relevant data sources, and train the model via an iterative cycle that takes the model from inception (or experimentation) to production.

Chapter 4 discusses this data science workflow and the many software tools, libraries, notebooks, and frameworks it uses. These tools help automate common tasks, such as selecting algorithms, exploring data sets, building ML models, and scaling the resulting predictive analytic applications to serve a broad base of users in production environments.

Data Science in the Cloud

Perhaps more than any other single technology, cloud computing has been responsible for helping the field of data science progress from hype to reality. These projects require immense amounts of data storage capacity and computing power. The cloud makes these resources inexpensive and abundant, enabling data science teams to store nearly limitless volumes of data at a progressively lower cost, and processing that data via powerful arrays of computers that can be scaled up and down at will.

AN EASIER PATH TO SUCCESS WITH DATA SCIENCE

The cloud has opened access to near-limitless, trouble-free capacity, allowing data scientists to focus on predictive insights rather than managing infrastructure. Benefits include:

- Centralized storage for all data
- Near-unlimited compute resources on a pay-per-use basis
- Scalable environments to move ML models quickly from test to production
- Easy adoption of ML tools, libraries, and frameworks

The cloud has made data storage and processing less expensive, and also has removed the burden of managing hardware and software infrastructure from busy IT teams. Today's cloud computing models allow companies of all sizes to rent compute and storage capacity in small increments and to instantly scale their infrastructure footprints as needed.

REMEMBER

Rather than buying and maintaining excessive capacity in on-premises data centers, in the cloud, you should pay only for the storage and processing resources you use. This is significant for data science workloads, which often fluctuate widely in computing and storage capacity.

Dealing with Volumes and Varieties of Data

For decades, businesses have been collecting data to streamline operations and inform decision-making. In years past, most of this data came from online transaction processing (OLTP) applications, which were deployed to manage finance, human resources, customer relationship management, supply chain, and other business functions. Much of the data was stored in relational databases, managed by in-house database administrators (DBAs), and maintained in corporate data centers. The data was staged in data warehouses, displayed through BI dashboards, and infused into everyday business processes.

Today, with the rise of countless mobile apps, pervasive social media networks, and billions of Internet-connected devices, making sense of organizational data requires solving problems on an entirely different scale. In its 2020 Global DataSphere forecast, IDC predicts that worldwide data creation and consumption will rise at a 26 percent compound growth rate between 2020 and 2024.


TIP

The best way to discover meaningful patterns in these disparate data sources is to consolidate the data into a data platform, which supports *structured* data in relational tables; *semi-structured* data from IoT devices, weblogs, and other sources; and *unstructured* data, such as image files and PDF documents.

Deploying a Cloud Data Platform as a Foundation

Popular cloud services, such as Amazon Web Services (AWS), Microsoft Azure, and Google Cloud, provide nearly boundless repositories for storing and analyzing data. However, they lack essential software services such as scalable security and governance to unify analytic workloads and streamline data science activities. To complicate matters, each of these public cloud services uses unique technologies for ingesting, storing, and managing data, making it difficult to migrate data and applications among them.

These limitations have given rise to modern cloud solutions that complement and extend these public cloud offerings. Chief among them are *cloud data platforms*.

A cloud data platform is a specialized cloud service optimized for storing, analyzing, and sharing large and diverse volumes of data for many types of analytic workloads. Other chapters in this book detail how such platforms can streamline every aspect of the data science workflow. For example, a cloud data platform eliminates the inconsistent results that arise when various work groups use different copies of the data. A cloud data platform also unifies data security and data governance activities by ensuring all users leverage a single copy of data (for more on this, see Chapter 2). Finally, a cloud data platform fosters collaboration and ensures the organization has a scalable data environment for data science and related analytic endeavors.

A cloud data platform extends your analytic horizons by allowing you to store your first-party data and leverage a network of data from second- and third-party data providers. Additionally, you can easily and securely share data sets, internally and externally, with authorized members of your data science ecosystem.

WHAT TO LOOK FOR IN A CLOUD DATA PLATFORM

TIP

A cloud data platform enables the following benefits for data science teams:

- Unlimited performance and scale, with a pay-as-you-go pricing model
- Near-zero maintenance, with all resources managed by the data cloud provider so there is no need to perform software updates, tune databases, or perform myriad other administrative tasks
- Access to structured, semi-structured, and unstructured data, both internal and from third parties via a data marketplace — all with consistent security and governance
- A global presence with centralized security controls that satisfy unique data locality and sovereignty requirements
- One collaborative platform for many workloads, including data lakes, data warehouses, data engineering, data science, data analytics, and data applications
- Support for every aspect of the data science workflow — from data exploration to model development to deployment of production models and business-ready analytics
- Support for many popular ML frameworks, tools, and languages, including SQL, Java, Scala, Python, and R

IN THIS CHAPTER

- Leveraging first-party, second-party, and third-party data
- Centralizing data governance
- Eliminating data silos
- Reviewing the advantages of a cloud data platform

Chapter 2
Increasing Access to Data While Reducing Risk

Today's organizations need to collect, unify, integrate, analyze, and share disparate data sources as a foundation for modern data science operations. Machine learning (ML) algorithms learn from data; the more data you provide, the more capable they become. But, how do you ensure that your business users and data scientists can easily access a secure, consistent, governed data set — without adding onerous constraints that inhibit innovation and increase time to insight? This chapter describes the immense and varied data landscape confronting today's data science teams and emphasizes the importance of having a clean, governed, single source of truth for data science initiatives.

Gathering Data for Analysis

Most organizations generate and collect data from their critical software applications and direct interactions with customers online and offline. To amass a complete data set, they may combine this first-party data with second-party data from partners

and suppliers, and third-party data from a service provider or data marketplace.

- **First-party data** is internal data produced via everyday business interactions with customers and prospects.
- **Second-party data** is produced by or in collaboration with trusted partners, such as product inventory data shared with an e-commerce or retail sales channel.
- **Third-party data** can be acquired from external sources to enrich internal data sets. Common examples include COVID-19 health data and financial market data.

For example, if an airline wants to understand the impact of weather on flight delays, it might combine its internal operational data with second-party data from travel agencies and third-party data from a weather-forecasting service.

Good data powers good models. In some cases, the data a company collects becomes more valuable than the company itself. *Forbes* highlighted this fact in a report titled "Your Company's Data May Be Worth More Than Your Company." In the wake of the downturn brought on by the COVID-19 pandemic, United Airlines and American Airlines secured multibillion-dollar loans by collateralizing their MileagePlus and AAdvantage customer-loyalty programs. Third-party appraisals of this customer-loyalty data were valued at three times more than the market value of the companies.

REMEMBER

Data scientists must look beyond the first-party data they own and consider second- and third-party data to increase the predictive performance of their models and discover new patterns, correlations, and insights.

Quantifying the Value in Data Science Endeavors

Amassing the right data is essential — and can have a significant impact on success. Respondents to a McKinsey global survey on artificial intelligence (AI) titled "The state of AI in 2020" revealed that a growing number of organizations are using AI as a tool for generating value — increasingly, in the form of revenues. AI leaders in a variety of industries attribute 20 percent or more of their organizations' earnings before interest and taxes (EBIT) to AI, McKinsey reported.

DATA ACCESS CHALLENGES

Succeeding with data science begins with getting your arms around all relevant data. To take stock of your particular situation, you must acknowledge and confront the following challenges:

- **Data is everywhere.** Tracking down all your data, and determining who is accessing it, can be problematic. Getting your arms around this problem begins with identifying, rationalizing, and eliminating data silos, both inside and outside your organization.
- **Governance is complex.** Knowing the kind of data you have — and how it's used — can be difficult, particularly when data sources vary in type, location, and rate of change. Centralizing data in a universal repository is the first step to reducing complexity.
- **Collaboration is complicated.** Data silos make it hard to capture and share data within data science workflows and complicate security and compliance tasks. Putting your data in a modern cloud data platform can simplify collaboration and reduce risk.

ML algorithms use statistical techniques, such as regression analysis and time-series modeling, to foresee pertinent developments and perform specific tasks. However, with the proliferation of open-source tools, libraries, and frameworks, many of these software assets have become commodities. The real value in today's data science initiatives lies beneath the surface: in the data sources, data models, data pipelines, and other software infrastructure that allow data science teams to capture, transform, and present data for processing.

Centralizing Data Governance

ML models can improve business processes and make knowledgeable decisions only if they have abundant data relating to the problems they're attempting to solve. But, how do you put the right data in the hands of the data scientists and business analysts who build these models without compromising data privacy or putting your brand at risk?

Data governance entails knowing precisely what data you have, where it resides, who is authorized to access it, and how each type

of user is permitted to use it. Instituting comprehensive controls reduces the risk of compliance violations. All data governance strategies should seek to protect sensitive data as it is accessed, shared, and exchanged across the organization and beyond. Ideally, these operations should be invisible to data scientists: They should protect the organization by ensuring their models meet the necessary standards without becoming onerous or burdensome.


TIP

To move data science projects from prototype to production without incurring data privacy or security violations, you need to set rigorous access policies and enforce consistent controls wherever data is used (see Figure 2-1). For example, all users must uphold data privacy laws that govern consumer data while also adhering to regional, national, and industry regulations.

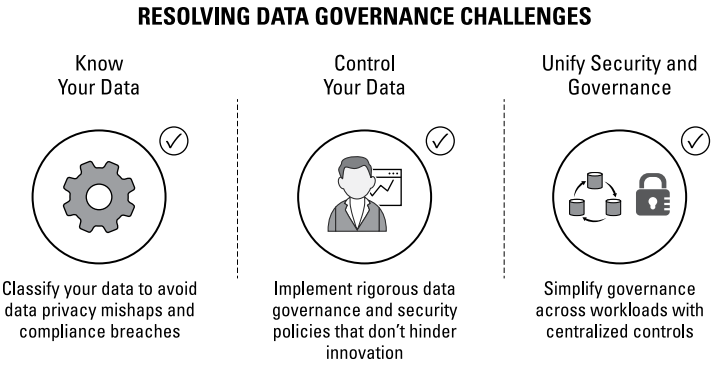

FIGURE 2-1: Comprehensive data governance is the bedrock of data science.

Drilling into data governance

At a granular level, appropriate security and governance must include object tagging, row-access policies, dynamic data masking, and external tokenization.

- **Object tagging** controls access to confidential and sensitive information such as salary amounts and Social Security numbers. For example, you might set permissions so that a line-of-business manager can see salary information for employees within their department, while an HR manager can see this information for all employees. For other people working with this data set, such as software developers and data scientists, those fields can be either partially or wholly masked. Applying and managing these tags should be easy

for the department that owns the data. Furthermore, the rules the tags define must apply universally whenever and wherever that data is used.

- **Row-access policies** allow users to see only the information relevant to them. With an umbrella policy to restrict access, users don't have to issue constraints in their queries, and security administrators can apply consistent access restrictions across all workloads. For example, you might implement a policy that ensures sales reps can see rows of customer data only for their accounts, while regional managers can see all the customer data from their regions.
- **Dynamic data masking** is a column-level security function that uses masking policies to selectively mask data at query time that was previously loaded in plain text. Some cloud platforms let you use data masking and data de-identification techniques to store personally identifiable information (PII) without making it visible to unauthorized users.
- **External tokenization** enables organizations to tokenize sensitive data with a third-party service before loading it into their cloud data platform. Tokenization turns the data into a random string of characters that has no meaningful value if the organization's system is breached. They can then dynamically detokenize the data at query runtime by using masking policies on functions stored and executed outside the platform. Once data is tokenized, visibility is very similar to dynamic data masking: If you're authorized, you see the clear data; if you're unauthorized, you see the tokenized data.

Democratizing data access

Data scientists cast a wide net as they gather data from internal and external sources to develop, train, and put predictive models into production. Data governance is often lacking, which is why data science initiatives can fail to scale beyond a small group of people.

REMEMBER

Without centralized data governance, data scientists are charged with identifying the correct data owners within each business unit and issuing multiple requests for the data they need, delaying projects and hindering collaboration among business units.

TIP

IMPROVING DATA VALUE

Data gains value the more it is used. Follow these principles to expand your data horizons:

- **Foster agility.** Access to new types of data can significantly impact how your business discovers new opportunities and responds to change.
- **Encourage innovation.** As new data sources become available, data science teams can explore new types of predictive analytics and apply their knowledge to improve operational efficiencies and drive new ways to improve customer experiences.
- **Expand your horizons.** Governed data has value beyond your internal organization and can also be monetized throughout your ecosystem of customers, partners, and other constituents.
- **Enrich your data foundation.** Expand your sphere of analysis and enrich analytic outcomes by integrating second- and third-party data with your internal data.

Isolating workgroups in this way also curtails one of the fundamental tenets of data democratization: the more people who can use and analyze the data, the more valuable it becomes. For example, a purchase order for raw materials may be visible only to the finance team that approved the job and perhaps the line-of-business manager who ordered the materials. Could it also be valuable to a brand manager attempting to identify upsell opportunities within the customer base? How about a shop-floor manager charged with monitoring material flows on the company's production lines?

Extending governance for all data types

As the sheer volume and variety of data continues to rise, many organizations struggle to support structured, semi-structured, and unstructured data in a cohesive, governed way. Cloud services, such as Amazon S3, Microsoft Azure Blob Storage, and Google Cloud Storage, make it relatively easy to store unstructured data, such as PDF files, audio, and video. However, granular access controls such as row-level permissions are not always available at the blob level. These services may broaden your access to complex data types, but not without increasing risk.

Loading all data into a centralized repository with a cohesive layer of data governance services allows you to enforce universal policies that broaden access while reducing risk. Applying these controls in a modern cloud data platform that supports structured data, semi-structured data, and unstructured data is easier and less risky.

REMEMBER

Consolidating your data in a cloud platform can simplify data governance activities. Ensure your platform vendor supports granular data access controls, such as object tagging, row-level access, and data masking.

Unifying Data Access with a Modern Architecture

Many analytic applications and machine learning (ML) models leverage unique, individualized data sets because analyzing data from multiple sources is difficult. These data silos almost always are accompanied by fragmented analytic systems and complex data pipelines that require custom coding and specialized extract, transform, and load (ETL) technology to capture data and prepare it for analysis.

In some cases, data science apps maintain data in a different form from business intelligence apps, each necessitating a separate copy of the data. These disparate data management practices make reconciling differences and synchronizing data among projects and teams complex. The inherent dissimilarities also foster data silos, which complicate governance by making it difficult to trace the data's lineage, catalog the data, and apply security rules (see Figure 2-2). These complications slow getting data ready for analysis, delay the time to insights, and raise the risk of ML models breaking down because of brittle data pipelines.

Tracing the rise of interim solutions

To make sense of all these types of data, some organizations built data lakes based on the Apache Hadoop ecosystem. As a processing engine for transforming large amounts of data, Hadoop was a breakthrough. It pioneered the concept of scaling analytic tasks linearly by combining inexpensive, commodity hardware systems rather than scaling vertically with large and expensive specialized computers. Hadoop also paved the way for the separation of

storage and compute resources — an architectural motif that persists in today's best cloud data platforms.

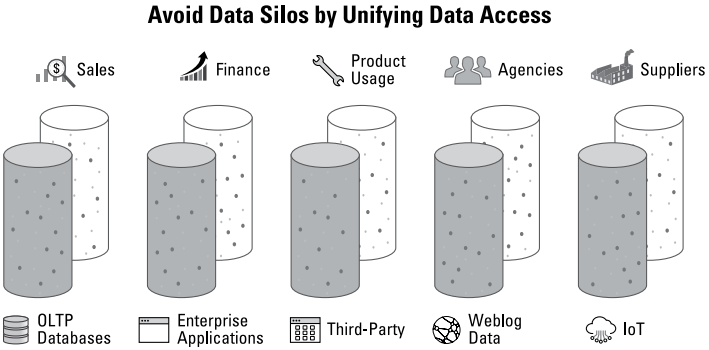

FIGURE 2-2: Without a centralized repository or cloud data platform, each analytic project creates a new data silo, complicating the data science lifecycle.

However, simply putting data into a data lake doesn't automatically lead to insights. Building, operating, and scaling data lakes is expensive and operationally complex because they require specialized skill sets to operate. Rudimentary management controls and minimal data security inhibit collaboration, resulting in new data silos — such as data warehouses for structured data, data lakes for semi-structured data, and object stores for unstructured data.

Consider the many ways data silos negatively impact data science initiatives:

- **Lost time:** Data scientists squander time unnecessarily attempting to identify, combine, and prepare data for analysis.
- **Lost value:** The longer an organization waits to act on its data, the more missed opportunities and the less overall value it generates.
- **Ineffective models:** ML models are only as good as the data they employ. A model that can't access all the data it needs will generate only partial value.
- **Skewed results:** Discrepancies in data among teams can lead to organizational misalignment and conflicting results in models, leading to inaccuracies in predictive applications, dashboards, and reports.

Leveraging a cloud data platform for data science

If your business analysts, data scientists, and other stakeholders can leverage a single source of truth across your entire organization, these teams can spend more time analyzing data and less time preparing it for analysis. Turn to Chapter 4 to see how establishing a common repository for all types of data and all types of workloads, including data engineering, in a cloud data platform empowers diverse teams to collaborate effectively throughout the data science workflow.

A cloud data platform can maximize your use of pertinent data in the following ways (see Figure 2-3):

- Enable data scientists to access multiple types of data without the complexity of managing infrastructure or copying and moving data
- Allow an extended ecosystem to securely share governed data, both internally among business units and externally with suppliers and partners
- Enhance existing data sets by accessing third-party data and data services via a secure, cloud-based marketplace

How a Cloud Data Platform Can Create Access to Data

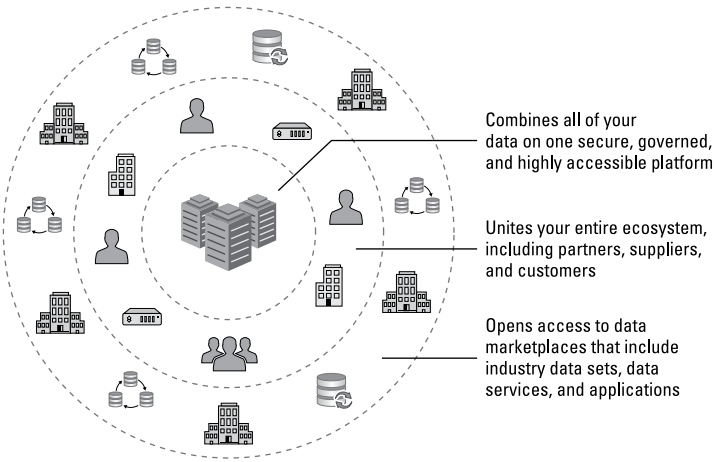

FIGURE 2-3: A cloud data platform serves as the central hub for data science endeavors by combining data, uniting the ecosystem, and opening access to external data sets and data services.

POWERING A LARGE E-COMMERCE OPERATION

CASE STUDY

Demand for Instacart's on-demand grocery delivery service surged during the COVID-19 pandemic, challenging data scientists to create sophisticated models to predict escalating demand.

With the organization's tens of thousands of retail partners, hundreds of thousands of personal shoppers, millions of customers, and tens of millions of SKUs, Instacart's data scientists crunch an enormous amount of data to segment customers, understand regional markets, and make personalized shopping recommendations. Every product team has data scientists, and every process is attached to some form of a dashboard that tracks and measures outcomes.

Instacart's data scientists combine inventory data, point-of-sale data, and third-party data on weather, events, COVID-19 infection rates, and other variables to continually balance supply and demand. Initially, they stored all this data in a "cloud-washed" data warehouse, which resided in the cloud but had not been architected to fully leverage the cloud's boundless resources: The data warehouse didn't scale well enough to handle burgeoning usage. The team moved their analytic operations to a cloud data platform to reduce complexity, simplify collaboration, and ensure better control. Today, this data platform is the company's master repository for planning, execution, and analytics as well as operational processes, such as invoicing.

Using the platform to store its e-commerce catalog enables Instacart to onboard new partners fast. Rather than building e-commerce platforms, retailers can join up with Instacart in a matter of days and upload their product data within hours. All data is transformed and normalized as it loads into the platform, resolving duplicate product entries and correcting price anomalies.

Instacart's software developers, data scientists, data engineers, and business analysts find the process of developing new ML algorithms simpler and more efficient with a cloud data platform. Because the platform uses ANSI-standard SQL, democratizing access for everyone is easy.

IN THIS CHAPTER

- Discussing key roles and responsibilities
- Reviewing staffing considerations
- Automating common development tasks
- Synchronizing team activities
- Improving collaboration with a cloud data platform

Chapter 3
Collaborating on a Data Science Team

The traditional data science workflow engages multiple team members of differing skill levels. Although the exact positions vary from company to company, the most common roles include data scientists, data engineers, data analysts, machine learning (ML) engineers, and product managers. This chapter examines the responsibilities of these team members within the context of the data science workflow and offers tips for improving collaboration and synchronizing team activities.

Introducing the Major Players

Data analysts formulate the business problems data scientists are asked to solve. These individuals know where data is located, they have experience using data visualization tools to explore possibilities and discover insights in the data, and they have the business domain expertise to explain their findings to the business community. They're skilled at defining metrics and key performance

indicators (KPIs) and at building dashboards and reports, using either SQL, a business intelligence (BI) platform (such as Tableau), or a programming language (such as R).

Data scientists see the business problem from a slightly different perspective. They use their math, statistics, and analytics skills to identify useful data sets and derive meaning from those data sets. These technology professionals use a combination of tools, applications, principles, and algorithms to make sense of the world's growing volume of data. They know how to develop artificial intelligence (AI) models to solve specific problems and are often familiar with Python, R, SQL, and data science notebooks.

Data engineers are charged with making clean and organized data available to the team. These technology professionals are familiar with testing and maintaining database management systems, creating data models, and using extract, transform, and load (ETL) technology to create data pipelines.

REMEMBER

Whereas data scientists develop and train models that make reliable predictions, data engineers are responsible for the data pipelines that feed these models. They're comfortable using Java, Scala, Python, SQL, and various ETL tools to create efficient data pipelines. They may also use Kubernetes, Terraform, Apache Kafka, or Airflow.

ML engineers know how to integrate ML models with existing infrastructure and data and scale new infrastructure as data volumes grow. They straddle the line between data science and software engineering. They're often skilled in probability and statistics and should be comfortable using Java, Python, cloud infrastructure, and possibly Terraform, Docker, and Kubernetes.

Product managers, and to an extent, executive sponsors, have a broad understanding of the issues and needs within their purview. They rely on this perspective to guide the development of solutions that meet the needs of the end-user community, and they help manage projects to make sure those solutions are implemented in a timely manner.

Starting Small, Growing Organically

According to the 2020 "Future of Jobs Report" from the World Economic Forum, the highest demand for strategic, skilled workers includes data analysts, data scientists, AI and ML specialists, robotics engineers, software and application developers, and digital transformation specialists.

With the shortage of talent, many companies help their staff members grow into new positions, which opens opportunities for IT professionals to transition to new roles as businesses become more data-driven. For example, as more applications and databases move to the cloud, software developers and database administrators can shift their focus from managing on-premises hardware and software to understanding the data's potential to move the business forward.

TIP

You may not need all these professionals at the outset, so focus initially on the roles mentioned in the "Essential Roles and Capabilities" sidebar. For small organizations that can't afford a large technology team, you may assign one individual two or more of these roles. For example, a data scientist might also take on some of the responsibilities of a data analyst.

ESSENTIAL ROLES AND CAPABILITIES

Large organizations may have significant data science teams that include all the positions discussed in this chapter, but these roles are crucial:

- **Data scientists** build ML models and train them with data. They use notebooks, such as Jupyter and Zeppelin, and languages, such as R, Python, Java, and Scala.
- **Data analysts** are the business experts. They know the data and are comfortable with tracking and reporting on metrics and KPIs.
- **Data engineers** prepare data and establish automated data pipelines that feed ML models continually.

Embracing ML Adoption across the Organization

Creating a data-driven culture that is skilled with ML and AI requires openness and collaboration. Data science should be integrated throughout the business to drive product improvements, improve operational efficiencies, and discover new revenue opportunities. As illustrated in Figure 3-1, data scientists are most successful when they work from the same single source of truth as the rest of the organization.

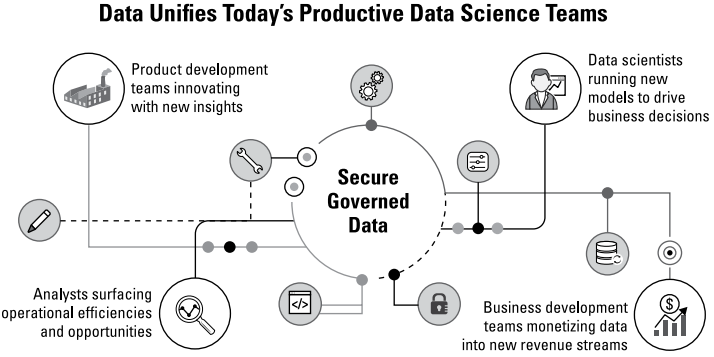

FIGURE 3-1: A secure and governed source of truth helps to increase adoption of ML-driven insights that drive new products and efficiencies.

Data science begins with experimental iteration. Data scientists identify a business problem, create a model, and then train that model to test their hypothesis. They collect and cleanse relevant data and then assess and gradually refine the business requirements as they visualize that data. Finally, if the experiment works as expected, they put the model into production.

TIP

Many teams instinctually try to improve the code when an algorithm isn't performing well. For many practical applications, however, focusing on improving the data is more effective.

To extend ML beyond the hands of a few experts, organizations can reduce the barriers to adoption by using automated ML tools, commonly called AutoML, which are designed to automate many steps in the data science lifecycle, empowering data analysts and business professionals to develop models that formerly required highly skilled data scientists.

TIP

Business experts with extensive domain knowledge and some AutoML training can sometimes take the place of experienced data scientists. Building a base of AutoML practitioners can help ease the skills shortage and create an able talent pool to complement highly trained data science resources.

According to a McKinsey survey titled "The state of AI in 2020," companies that lead their industries in the use of AI technology are twice as likely as their counterparts to use standardized, end-to-end platforms for AI-related data science, data engineering, and application development. Rather than having various teams proceed in isolation from each other and amassing many different tools — each with its own copy of data — these companies acquire technologies that leverage a single source of truth to encourage collaboration.

Improving Collaboration with a Cloud Data Platform

How do you build a holistic picture of your business to find strategic insights when no single application contains all the information you need? Consider using a cloud data platform — a specialized cloud service optimized for storing, analyzing, and sharing large and diverse volumes of data and for putting that data to work in the context of many types of analytic workloads.

For example, a sales team may want to score accounts, which involves ingesting data from internal customer relationship management (CRM), enterprise resource planning (ERP), and sales force automation (SFA) systems. This enterprise data may be enriched with third-party account firmographic and technographic data, cleansed to eliminate duplicates and low-quality data, and finally fed into a scoring model to gauge each account's propensity to buy.

A cloud data platform anchoring these endeavors can combine these various data sources according to good data governance practices (see Chapter 2) and ensure fluidity among data science, analytics, and data engineering workloads. A cloud data platform can also serve as the control center for sharing data among key business applications, such as connecting customer data in Salesforce with vendor data in Workday. Once these data sources are centrally located, you have a single source of truth for all departments, from sales to finance and procurement.

Here's another way to look at it: A cloud data platform minimizes the amount of code between you and your data. Because some platforms support structured data, semi-structured data, *and* some forms of unstructured data, you can use a cloud data platform for your data lake and your data warehouse, bringing the two together. For data scientists, this dramatically accelerates the creation of ML models by making it easier to access all forms of data and helps harmonize related analytic activities. For example, the results from data science experiments can be pulled back into the platform, where they can be exposed through reports and dashboards.

REMEMBER

By bringing all your data together, you eliminate barriers among databases, applications, and business functions — no matter where the data comes from or how it is structured.

SYNCHRONIZING TEAM ACTIVITIES

Standardizing on a cloud data platform for data science projects allows all members of a diverse team to work in concert, while giving team members the flexibility to work with their most productive tools and languages. You can empower multiple groups to access the same data simultaneously without impacting performance or requiring data movement, data marts, or copies of data. Additionally, you can automatically scale analytic endeavors to support a near-unlimited amount of data, workloads, and concurrent users.

Other benefits of a modern cloud data platform include the following:

- Multicluster compute enables scalable data processing — using popular data science languages including SQL, Python, Java, and Scala — for any number of concurrent users and workloads.
- Per-second compute pricing and affordable cloud rates for storage provide a cost-efficient solution for data science initiatives.
- Near-zero maintenance simplifies data administration chores and allows workers to focus on model development instead of systems management.
- Data marketplaces and exchanges expand potential data sources to enhance model performance.
- Broad support for libraries, notebooks, frameworks, and tools accelerates the development of predictive models and data science apps.

IN THIS CHAPTER

- Exploring data science processes, tools, and methods
- Identifying pertinent data sources
- Extracting value from your data
- Using data science tools to automate common tasks
- Circumventing problems facing data science teams

Chapter 4
Facilitating the Data Science Workflow

Successful data science initiatives depend on getting the right data at the right time to the correct models. Data is added and prepared multiple times during each stage of the data science lifecycle, often with different data requirements. This chapter walks you through the major steps in this workflow and emphasizes the importance of solid, consistent data management practices.

Defining the Business Problem

Before diving into a new project, consider these three questions:

- Are the problems and objectives well defined?
- Is there an executive sponsor for the project?
- Does the project have a clear business impact, and will it yield a measurable payback?

Next, determine whether artificial intelligence (AI) and machine learning (ML) are required to obtain the answers you seek. If you

wish to extract historical insights from data, you can use basic analytics and business intelligence (BI) tools.

Outlining the ML Process

Data science follows a fundamentally different workflow than the more familiar software development lifecycle, which focuses on implementing well-defined functional requirements. Rather, the data science workflow starts with a business objective, and the requirements are uncovered gradually via an iterative process of data exploration and model development. As shown in Figure 4-1, the six stages in this process are broadly divided into two phases: a *prototype phase* (stages 1 through 4 in the workflow) and a *production phase* (stages 5 and 6).

The Six Stages of the Data Science Workflow

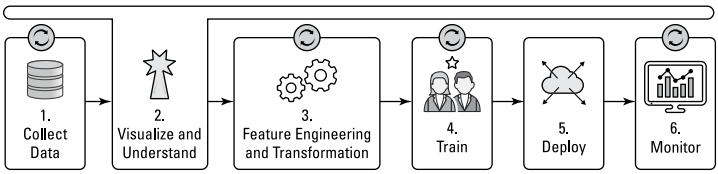

FIGURE 4-1: Understanding the data is pervasive throughout the process.

Collecting data

It's easy to say that data science is the easy part. Getting the right data and preparing it for analysis is much more difficult. These steps often consume the majority of the time spent within the data science workflow — especially if your data is scattered across multiple systems. Collecting and combining data is important. But it can take valuable time away from other essential data science work unless you have a modern cloud data platform that supports all data types, including semi-structured and unstructured data. It should provide one place to instantly access all relevant data in order to get the most predictive power out of ML models.

With secure, governed data at the heart of your data science workflow, you can democratize that data to reveal new insights across your organization. You can also easily and securely share second-party data with partners and suppliers, and share third-party data via a data marketplace or data exchange, to reveal previously unimagined insights.

Visualizing and understanding data

After data scientists define the business objectives, they embark on a prototyping process. This involves multiple iterations of visualizing and exploring the data. The goal is to transform the collected data into *features* — attributes derived from the data — as inputs into models for training and evaluation. They experiment with many different data sets throughout this iterative process, and they may use BI tools or interactive notebooks to better understand the data, visualize it, and formulate hypotheses.

During this stage, data scientists focus on understanding the *quality* of the data they have selected. What is the distribution of the data? Are there outliers that need to be addressed? How complete is each column? Are important data values missing? Data cleansing may be necessary to improve the data set and ensure accurate predictions. For example, if your model is designed to predict why some e-commerce customers abandon their shopping carts while others click through to checkout, you need complete and consistent data on web browsing activities and clickstream patterns.

Preparing data for model training

Feature engineering is the process of using domain knowledge and algorithms to generate new fields from the raw data, such as extracting the hour of the day from a time-stamp field to track order fulfillment performance. Many models accept features only in numeric form, requiring categorical data coding schemes and other transformations to reduce the number of distinct values.

Performing feature engineering on data located within a shared, centralized, cloud-built platform can accelerate model development, reduce costs, and enable the reuse of new features. Central access to features helps maintain consistent, governed data for training, evaluation, and aiding reproducibility (the ability to obtain the same results when replicating an experiment or procedure).

Training ML models

After a feature set is in place, data scientists "train" the models, which involves passing a subset of the data, or *training set*, through the modeling function to derive a predictive function (trained model). Each model is tested on data that wasn't part of the training set to determine its effectiveness — usually measured in terms of its predictive strength and robustness.

The technology landscape for model training and development includes analytics tools, libraries, notebooks, and programming languages. The combination of these enables data scientists to conduct in-depth feature engineering, tune hyperparameters, and iteratively create and assess ML models. Intuitions become accurate predictions as data scientists experiment with various algorithms, assess the performance of each model, and adjust parameters on their chosen model to optimize the code for their specific data sets.

- To prepare data and create ML logic, data scientists use programming languages such as Scala, Python, R, and SQL.
- To build, refine, and train the models, data scientists use ML frameworks and libraries. Many of them are open source. Popular options include scikit-learn, XGBoost, LightGBM, TensorFlow, Keras, and PyTorch.
- To refine the models and collaborate with coworkers, data scientists use ML notebooks, such as Jupyter. These notebooks are available as adjuncts to leading data science tools, including Amazon SageMaker and Dataiku.

Data analysts and other users who want to benefit from ML without having advanced programming skills or deep knowledge of mathematics and statistics can take advantage of AutoML tools, such as RapidMiner, BigSquid, H20.ai, and DataRobot. These tools make it easier to select algorithms, train models, and ultimately choose the best model for the business problem at hand.

Identifying the right ML models

According to Analytics Vidhya, which provides a community-based knowledge portal for analytics and data science professionals, most ML models fall into three categories:

- **Supervised learning algorithms** include a target or outcome variable predicted from a given set of independent variables. Examples include regression, decision tree, random forest, k-nearest neighbors (KNN), and logistic regression models.
- **Unsupervised learning algorithms** are used for population clustering models, such as segmenting customers in different groups for specific campaigns or understanding what items are commonly sold together. Examples of unsupervised learning algorithms include Apriori and k-means.

» **Reinforcement learning algorithms** make specific decisions by using trial-and-error training processes. These algorithms learn from experience as they capture the knowledge to make accurate business decisions. Reinforcement models typically are modeled using a Markov decision process and use deep learning neural networks in training.


TECHNICAL STUFF

When selecting models, data scientists must also balance *accuracy* and *explainability*. Accuracy refers to the capability of a given ML function to predict a response value for a given observation. Explainability refers to visibility into a model's decision-making methods.

For example, in a predictive maintenance scenario, an alert needs to be accompanied by the variable or set of variables triggering the predictive alert — whether it's temperature, vibration, or some other variable. *Linear regression* is a common example of an algorithm that can explain the relationship between the model input variables and the prediction. On the other hand, other models leverage complex functions, providing little or no explanation of how each variable impacts the model's overall predictive outcome. This makes it difficult to determine how the model draws its conclusions.

Deploying and monitoring models with MLOps procedures

After the model is selected, it needs to be deployed — meaning integrated into the production data pipeline or application. The data transformations applied to the data in the training data set must be replicated identically and at scale on the new data from which predictions are sought. To do this, data engineers and ML engineers engage with data scientists to refine the model for a live business context.

MLOps — short for machine learning operations — is a set of best practices for collaboration and communication among data scientists and operations professionals. Synthesizing ML, DevOps, and data engineering, MLOps helps structure the data science workflow and ensure stable change management processes as projects move from inception to production. MLOps procedures ease maintaining consistent data pipelines and managing data versions as models are trained, tested, validated, and placed in production. They also help operations professionals manage updates and detect model deterioration. The goal is to provide end-to-end reproducibility, governance, and lineage throughout the ML lifecycle.

ML engineers monitor the models to check for *data drift*, meaning the input data is changing or the model is losing its predictive power. When drift is detected, the engineers commonly send the models through further iterations of retraining, feature engineering, and tuning, ultimately replacing the current best model in production with the improved model.

Addressing Common Barriers

Data science can be an expensive endeavor, so projects need to show value; therefore, stakeholders need to address common problems before they affect outcomes.

Historically, data science models fail to reach production for a few familiar reasons:

- Lack of focus on business value
- Security and privacy concerns, failure to uphold IT security and data privacy standards, or fear of information leakage via the model
- Complexity of integrating with existing infrastructure and IT resources
- Data problems, such as unwieldy amounts of data, overly complex data, or the inability to source data required for accurate prediction due to latencies in upstream processes and systems
- Complex systems that require manual operations and can't scale to meet data processing demands of production data pipelines
- Inability to automate and scale the model monitoring, validation, and retraining process to support additional production models

REMEMBER

Particularly at smaller firms with limited resources, data scientists may work solo and wear many hats. Although these one-person teams may be incredibly self-sufficient, their efforts don't always scale. Ad hoc data pulls; inconsistent data quality procedures; and isolated sandbox environments for training, validating, and testing models may not adhere to IT policies or comply with data security requirements.

LAYING THE GROUNDWORK

Consider the following guidelines when preparing for a data science project:

- **Defining success:** All stakeholders should develop consensus on the desired outcomes — and, ideally, develop key performance indicators (KPIs) and other measures to verify the endeavor's success and the predictive models' accuracy.
- **Collecting and cleaning:** Data scientists spend a lot of time gathering data and ensuring data quality. Involve data engineers to ensure that each potential data set will be available and properly prepared for ongoing consumption.
- **Feature engineering:** Extracting features from data is much easier with reusable code and a shared data repository. The closer you can co-locate the code and the data, the more process efficiency and data governance control you will have. Wherever possible, store and catalog features centrally to help data scientists discover and reuse rather than reinvent.
- **Model training:** Your code-savvy data scientists may use notebook-based tools. Your business experts may use AutoML tools. Either way, enable all users to integrate with your data source of truth to help ensure collaboration rather than conflict about who is using the right data inputs in their model.
- **Production:** Once a model reaches production, it is tasked with model inference — the process of running predictions on new data to infer associations and reach conclusions. Ideally, this action should be performed directly where the data resides to reduce the governance and security complexity associated with integrating models with production data.

Streamlining the Workflow with a Cloud Data Platform

A cloud data platform can streamline the entire data science lifecycle of ML, AI, and predictive application development. By leveraging a centralized, cloud-built data platform that provides near-unlimited data storage and compute power, data scientists can unite their siloed data, easily discover and securely share governed data, and execute data processing for all analytic workloads.

Data scientists can acquire the data they need to develop new applications quickly and train the associated ML models. They can run complex ML algorithms in the cloud and integrate external ML services there, such as invoking a risk-scoring model written in Python, R, or Scala. Data scientists and engineers can run these procedures externally or directly in the centralized data repository, making a cloud data platform ideal for data preparation and feature engineering. Analysts can pull in the result set (predictions) for model monitoring and surface the insights via BI apps, which also leverage the cloud data platform. Figure 4-2 illustrates this convergence of data analytics, data science, and data engineering.

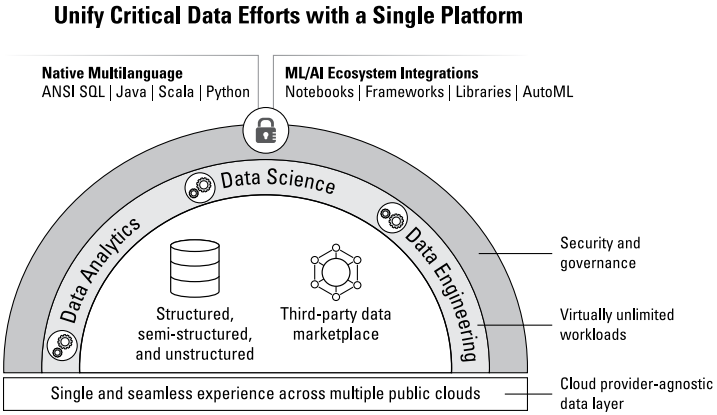

FIGURE 4-2: A cloud data platform that supports many data types and spans public clouds can unite the work of data analysts, scientists, and engineers.

Follow these guidelines to improve your chances of success with data science endeavors:

- **Reduce friction** by using a data-centric approach that lets you access nearly all data types, find data gaps, and enhance in-house data with second- and third-party data.
- **Boost productivity** by establishing a data foundation that encourages freedom of choice: any framework, any tool, any language, and any cloud provider.
- **Enhance collaboration** with a cloud data platform that enables a single source of data and a near-unlimited number of concurrent and diverse data workloads.

IN THIS CHAPTER

- Personalizing customer experiences to boost engagement and grow sales
- Optimizing operations to reduce operational costs
- Detecting anomalies in business data to mitigate risks

Chapter 5
Applying Data Science to Your Business

This chapter discusses common horizontal use cases that span many different industries, including manufacturing, retail, energy production, insurance, and financial services.

Personalizing Experiences

Nearly everybody has enjoyed personalized experiences from companies that make an effort to get to know their customers' profiles, problems, and preferences. Pioneered by digital leaders, such as Amazon and Netflix, the ability to offer knowledgeable, timely interactions is now expected whenever consumers interact with vendors and other service providers.

Predictive personalization, driven by machine learning, enables businesses to build profiles of customers and present them with unique offers and services. Some firms conduct automatic real-time analyses to provide custom interactions, such as a clothing retailer that remembers your preferred styles, exact sizes, and favorite colors. Or a music service that curates personalized playlists based on your likes, dislikes, and listening habits.

CONSOLIDATING DATA FOR COMPANY-WIDE ANALYTICS

CASE STUDY

Pacific Life helps millions of individuals and families with their financial needs via a wide range of life insurance products, annuities, and mutual funds. Previously, the insurance giant used an on-premises data warehouse, but accessing 20 years of historical data wasn't easy. Some large queries would never complete unless the warehouse was segmented into smaller, more manageable slices of data. Additionally, the organization needed ten data engineers to develop and manage extract, transform, and load (ETL) processes.

Moving Pacific Life's data to a cloud data platform eliminated these constraints, enabling the organization's data science team to more efficiently analyze engagement patterns, equity returns, interest rate returns, and customer activity. They can now work with massive data sets, including third-party Experian data from an associated data marketplace, without manually copying, moving, or refreshing the data. Queries that took more than an hour to run on premises now run in less than three seconds. With near-real time processing of relevant internal and third-party data, Pacific Life can more accurately segment customers to provide more-tailored and personalized experiences.

Instead of creating silos of information, Pacific Life's entire business community can access a unified source of governed data, including historical policyholder data, presented via business intelligence (BI) dashboards. Data scientists and analysts across the organization can fully leverage their data, easily create workspaces to run their analytics, and then publish the resulting data sets for others to use. For example, data scientists can build models to predict policy-holder demand and surface results in business dashboards to optimize staffing schedules.

Satisfying customers by suggesting exactly what they prefer involves collecting purchase-history data, financial data, behavioral data, Internet-browsing data, and third-party data sets from data marketplaces and exchanges. The goal is to segment users, understand common behaviors and trends, and then apply general knowledge to specific users and groups. For example, by analyzing weblogs, you can monitor which pages and subpages customers visit, determine what they're looking for, and note where they

spend most of their time. You can get to know their habits from their clickstream activity, their biases from social media sentiment, and their tendencies from publicly available data.

REMEMBER

The richer each customer profile becomes, the better you can predict how individuals will respond to unique content and offers. The goal is to create stronger customer engagement and boost sales by offering customers something they actually want or are considering buying, rather than making them a random offer.

Artificial intelligence (AI) technology has become central to these endeavors. Machine learning (ML) algorithms can populate web pages with unique content for each shopper and even dynamically adjust prices to reflect competitive market trends. These intelligent systems can also segment customers based on their likelihood to purchase specific types of products. They can predict when people will return to your site and how likely they are to churn. Additionally, they can help marketers fine-tune their advertising campaigns to boost conversions and optimize revenue.

Optimizing Operations and Reducing Costs

Automated inventory management, just-in-time fulfillment, and other supply chain strategies have been important operational goals for decades. Today, verifying the performance of these operations has become a prerequisite for working with large retailers. Consumer packaged goods (CPG) companies, for example, must verify these performance metrics if they wish to participate in the Walmart Marketplace. According to Walmart's website, all sellers must uphold strict quality service metrics, such as a 90-day order defect rate of less than 2 percent, an on-time shipment rate exceeding 99 percent, and a valid tracking rate of at least 99 percent. An intelligent supply chain helps CPG companies and their supply chain partners meet these requirements.

For example, ML technology helps grocery store planners predict how many of each item they should stock in each store based on recent sales and projected trends, factoring in weather, public events, and various types of consumer data to ensure consistent delivery in both the physical and digital worlds. Advanced forecasting tools help planners anticipate sales volume, predict inventory needs, and maximize profit margins. In some cases,

they can also enable such things as dynamic pricing and stock reallocation between warehouses and distribution centers.

REMEMBER

Having an intelligent supply chain becomes progressively more critical as operations scale. Consider a large grocery store chain that includes thousands of stores and tens of thousands of products. These slim-margin businesses depend on AI to monitor sudden shifts in the flow of goods, such as when a global pandemic causes erratic consumer behavior or leads to supply bottlenecks. Sharing knowledge with other supply chain members can circumvent inventory shortfalls and help clear unused stock.

ENERGY COMPANY POWERS COMMODITY TRADING WITH AI

CASE STUDY

Uniper SE generates, trades, and markets energy across Europe and North America. The company's 34 gigawatts of installed generation capacity makes it one of Europe's largest power generators.

Uniper's market analytics team of 40 analysts and data scientists analyzes large amounts of data from more than 270 sources to provide insights that support commodity traders and guide investment decisions under ever-changing market conditions. Previously, ingesting data into an on-premises data warehouse took weeks of planning and administrative effort. Limited support for semi-structured data, combined with unreliable query performance, prevented data scientists from exploring high-volume data sets containing billions of time-series events.

Seeking to democratize data analytics and build a centralized platform to power its ML models, Uniper adopted a cloud data platform integrated with Microsoft Azure via a native connector for Python, among other tools. The platform democratizes data analytics for 200 users across multiple departments.

Having a centralized repository of structured and semi-structured data has increased the precision of existing models and allowed the market analytics team to develop 25 percent more models with the same number of personnel. Combining 250,000 time-series events in a single place and making them searchable and easily accessible fuels internal inquiry and enables cross-functional collaboration. Contention-free data exploration allows the market analytics team to transform raw data into meaningful insights twice as quickly and with less effort.

Retailers and CPG firms aren't alone. From energy production to life sciences, technology to communications, companies in many industries rely on AI technology to automate supply chains and streamline operational processes — such as using ML to automate commodities trading processes the way Uniper has done. Manufacturers share data with raw-materials suppliers to create accurate forecasts and fully leverage existing inventory. ML models also can provide insights to help coordinate sales, production, warehousing, distribution, maintenance, and stock-keeping activities. AI technology helps businesses reduce waste via more-efficient operations while boosting revenues and satisfying the needs of consumers, business partners, and other constituents.

Detecting Fraud and Processing Anomalies

Many companies use predictive analytics to detect anomalous behavior that might indicate fraud and to keep their skilled workers focused on the right customers, cases, and claims. For example, a property and casualty (P&C) insurer might create an ML model to detect *claims leakage* — dollars lost through claims-management inefficiencies and failures in business processes. Minimizing these leaks can dramatically reduce costs and improve the loss ratio, an important metric on P&C balance sheets. According to a 2019 PwC article called "Stopping the Leaks," addressing claims leakage can lead to savings of between 5 and 10 percent. For an insurer spending $500 million in annual claims, that translates to a bottom-line improvement of $25 million to $50 million — which could lead to lower rates for customers.

Claims leakage can result from procedural problems (such as inefficient claims reviews), insufficient claims monitoring, poor decision-making, or outright fraud. ML models can be trained to detect patterns in claims data to identify claims that have the potential for high leakage and enforce good business practices that increase claims settlement payments. They can also identify the trends that lead to excessive payouts, such as determining which adjuster should process each type of claim.

ML algorithms study thousands of claims to make knowledgeable recommendations, such as whether a claims processor should send an adjuster or tow a damaged vehicle to an auto body shop for inspection. A predictive model can calculate the loss associated with each scenario and suggest the best avenue to follow for each type of claim, augmenting the abilities of claim adjusters themselves.

TIP

You can apply these same risk dynamics to security analytics, preventive maintenance, credit card fraud, and anomaly detection in many industries, including cybersecurity, telecommunications, smart city planning, and energy research.

MITIGATING RISK FOR AN ONLINE LENDING FIRM

CASE STUDY

As one of Australasia's leading marketplace lending platforms, Harmoney has matched more than 50,000 borrowers with more than $1.6 billion in personal loans. The lending experience is 100 percent online, from application to funding, necessitating complete and accurate data collection, processing, and analysis.

According to the chief data scientist at Harmoney, AI is all about understanding the customer experience and studying the various steps in the customer journey. Harmoney uses ML models at loan origination to predict whether a customer is a low or high risk and appropriately price each loan. These models also assess the likelihood of loan defaults so the organization won't take on exceptionally risky clients. Harmoney maintains that using AI for these tasks is fairer and more reliable because it removes the possible misinterpretations and unconscious biases of human decision-making. Machines crunch the numbers, while employees focus on the tasks that require human knowledge and discernment, such as designing better customer experiences.

Maintaining data in a cloud data platform facilitates these crucial processes. Harmoney can store structured and semi-structured data in a single location, helping data scientists explore and interrogate all data from one unified repository. The data science team uses an AutoML tool to prepare, build, deploy, monitor, and maintain powerful risk-scoring applications at enterprise scale. The stakes are high: A 1 percent improvement in risk assessment delivers a $1 million gain to the bottom line.

IN THIS CHAPTER

- Building a solid data foundation
- Identifying pressing business problems
- Forming a skilled team
- Collaborating in a productive environment
- Verifying progress, meeting business goals
- Scaling data science activities

Chapter 6
Six Steps to Cloud Data Science

Data science is a journey that advances organizations from using data to describe what happened in the past to predicting and enhancing the future. Follow these steps to ensure positive outcomes:

- **Step one: Build a data foundation.** What data do you have today? What data will you need to solve tomorrow's business problems? Take advantage of a cloud data platform that supports multiple types of data captured from various types of devices and applications. Doing so will allow you to break down silos and extend consistent governance and security to disparate sources — internally from your enterprise applications and externally from business partners and third-party data providers. The platform should support popular data science programming languages, tools, and open source environments to maximize options for your team.
- **Step two: Identify the business problems.** Do you want to tackle analytic business problems or operational business problems? Perhaps you have some great business intelligence (BI) dashboards that show you *what* happened, such as an unexpected revenue shortfall. However, you've had trouble figuring out *why* these problems occurred, and you aren't sure

how to use analytics to improve the situation. If you want to predict an outcome, determine what will happen next, or make an educated guess about how a situation will evolve, you may need to build a machine learning (ML) model. Rank potential projects based on expected business impact, data readiness, and level of executive sponsorship.

- **Step three: Establish a skilled team.** What kind of team do you need to succeed with data science? At a minimum, you will need a data scientist or business analyst with the skills to build and train statistical models, a data engineer with experience building data pipelines and moving models into production, and a line-of-business leader or project manager to guide the effort. Before hiring new talent, see if you can train your existing team members to learn modern data science tools and adopt a predictive mindset.
- **Step four: Build a culture of collaboration.** How do you unify the efforts of various team members, each of whom may be pursuing unique data science projects? Standardizing on a modern cloud data platform enables everybody to access the same data simultaneously, without having to copy or move the data. Appoint a data steward to take charge of each data set — ideally a business owner who understands the data. Establish consistent procedures for data security, data privacy, and data governance to satisfy industry regulations and avoid compliance violations.
- **Step five: Measure, learn, and celebrate success.** How do you measure the success of data science initiatives? Start small, identify metrics to demonstrate business results, and validate progress with executive sponsors and stakeholders. Don't get derailed by temporary setbacks. If you don't obtain the results you were hoping for, step back, assess what went wrong, and try something else based on the lessons you learned. Apply successful outcomes to other departments and business problems.
- **Step six: Scale the effort.** As you move from small sandbox environments to production business applications, look to the cloud and its boundless data storage and compute resources. You can start small and expand gradually to scale the effort on a pay-as-you-go basis. Rather than pursuing multiple proofs-of-concept in isolation, share best practices and encourage reusability. Enhance your data sets with data from a modern data marketplace. Strive to democratize analytics and extend ML capabilities to the entire organization.

Predict and influence your organization's future with cloud data science

Data scientists spend nearly two-thirds of their time collecting, preparing, and visualizing data — leaving little time to build the predictive models that put that data to work. Data is inaccessible, computing bottlenecks stifle data science initiatives, and establishing a single copy of your data for diverse workloads to operate concurrently without performance issues is out of reach. Read on to learn how your organization can address these issues and establish a cloud data science practice to create intelligent and data-driven decision-making solutions.

Inside...

- Enable a modern cloud data platform
- Unify data silos and data governance
- Create a rock star data science team
- Define modern data science workflows
- Reveal new business opportunities
- Scale your data science initiatives
- Read real-world case studies

David Baum (david@dbaumcomm.com) is a freelance business writer specializing in science and technology.

Go to Dummies.com™
for videos, step-by-step photos, how-to articles, or to shop!

ISBN: 978-1-119-82939-3
Not For Resale

for dummies®
A Wiley Brand

WILEY END USER LICENSE AGREEMENT

Go to www.wiley.com/go/eula to access Wiley’s ebook EULA.