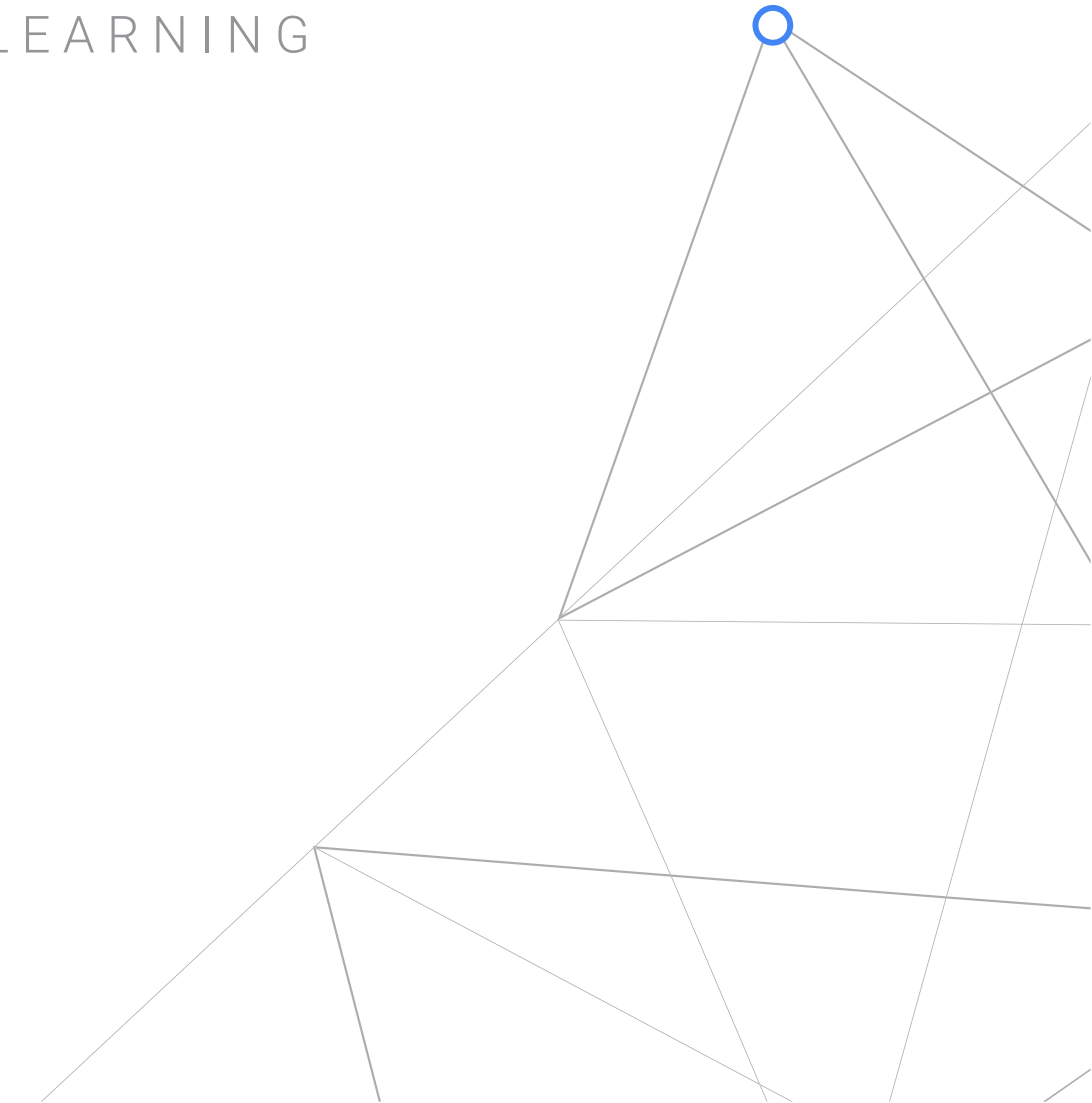
CIO'S GUIDE TO DATA ANALYTICS & MACHINE LEARNING



Google Cloud

CONTENTS

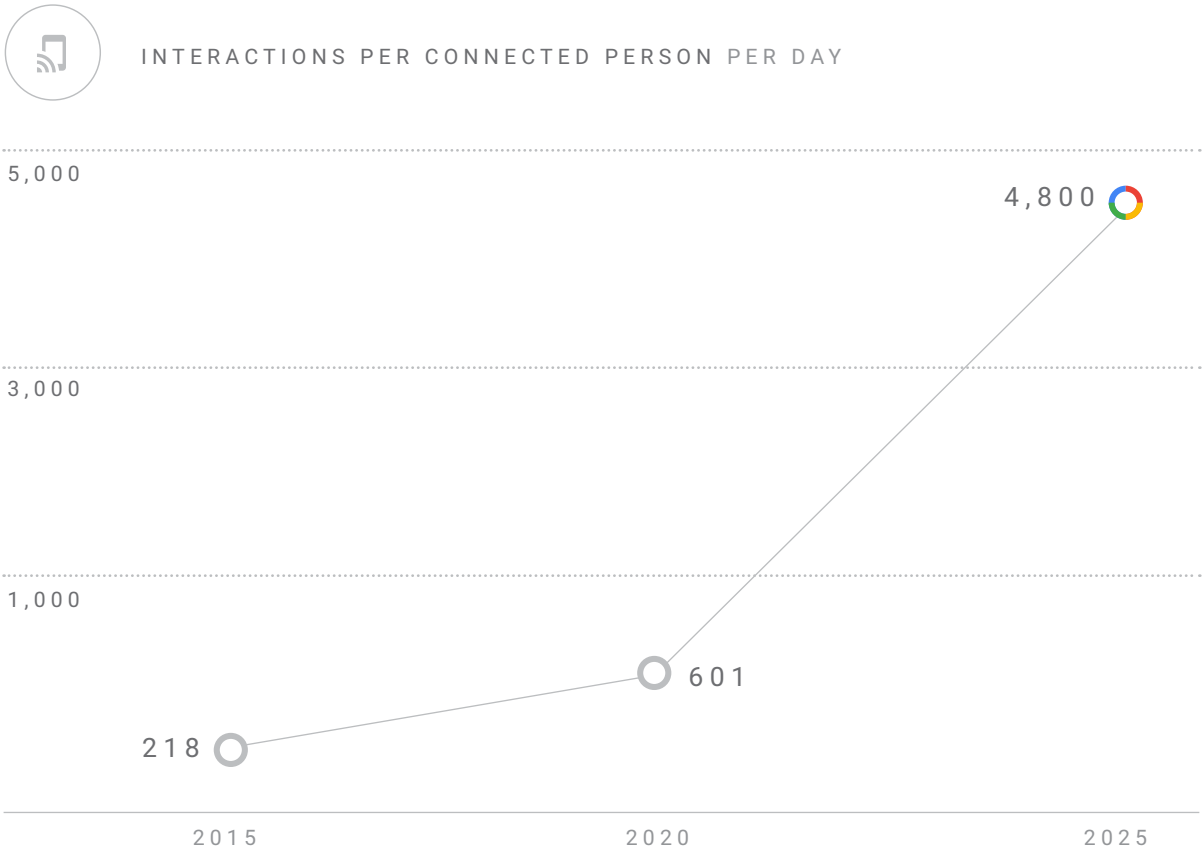
INTRODUCTION

Using data to make business decisions is nothing new. Once, "data-based decision-making" might have meant noticing a correlation between a print ad campaign and anecdotal accounts of higher-than-usual sales. Businesses used whatever data they could get, when they could get it.

Today, data is everywhere. It streams in from connected devices at dizzying speeds, in an array of formats, from billions of users. Big Data is often cast as an opportunity, but only for businesses that are structured to handle its volume and diversity. For other companies, the flood of data can represent a risk—that potential insights go untapped, customer needs go unmet, and businesses keep making uninformed decisions.

Two factors make the current landscape different from past evolutions. The first is an *exponential increase in the volume and diversity of data being generated* by billions of users and devices. The second is a *demand for immediate access to high-quality data and insights.* Each has brought new urgency to how companies manage data. In addition, the cost and performance of many cloud capabilities have reached a tipping point, helping make machine learning (ML) and artificial intelligence (AI) accessible to every business.

Despite widespread recognition of the value of data, few companies have implemented modern data strategies.[1] Building on original research and Google's own contributions in the cloud, this guide is designed to help IT and business leaders implement modern, cloud-based strategies for data management. In each section, we highlight technologies helping companies turn a vast, complex data landscape into useful business insights.



By 2025, the average connected person will interact with connected devices nearly 4,800 times per day—equivalent to one interaction every 18 seconds.[2]

OUR ROOTS

Google Cloud's *Guide to Data Analytics & Machine Learning* draws upon Google's twenty years of tackling some of the industry's toughest data problems. Along the way, we've contributed original research that has helped to shape the Big Data landscape: from two research papers in late 2003 and 2004, which together spawned the Hadoop movement; to the Dremel paper, which forms the basis for the cloud data warehouse capability you'll read about in this guide.

We designed, built, and deployed Spanner, the first system to distribute data at global scale and support externally consistent distributed transactions—and, in 2017, made it generally available to our customers.[3] More recently, Google Brain has helped fuel the industry's renewed interest in AI, leading up to the release of our TensorFlow Project into open source.[4] With this Guide, we look forward to sharing our experience with leaders looking for ways to unlock the promise of machine learning and AI for their organizations.

CHAPTER 1

THE NEW DATA LANDSCAPE

01

THE NEW DATA LANDSCAPE

01

Managing data would be easier if growth were limited to a few sources, or if data were uniform. The challenge lies in the diversity of sources and formats. This includes the growing volume of unstructured data: emails, system logs, web pages, customer transcripts, documents, slides, informal chats, and an exploding volume of rich media like HD images and video. Enormous volumes of information are available instantaneously from any device connected to the Internet, driving new expectations around the availability and immediacy of data.

Consumer applications like search, messaging, e-commerce, social networking, and online video were the first to encounter this problem. New systems had to be built to handle traffic at the scale of the entire Web, while returning insights instantly. These breakthroughs are now available—and increasingly important—to every business, from helping manufacturers manage supply chains more efficiently to increasing the precision of medical diagnoses.

IT teams are stuck in the middle. They must find ways to deliver a *real-time view of the business* while also managing a *larger and more complex data landscape*. As with many software initiatives, reducing complexity is an important determinant of success.

This Guide explores how managed cloud services help new and established companies meet today's data challenges. It presents a path that begins with capturing raw business data into cloud storage. As business questions arise, cloud-based tools can prepare and structure raw

data on demand. The prepared data is then integrated into a cloud data warehouse, where it's immediately available for analysis. This trove of data serves as a "home base" from which organizations can capture, prepare, and analyze data of any kind, from any source. The fully managed nature of cloud services helps streamline this entire process—including support for real-time analytics—without requiring IT to be aware of underlying infrastructure. Building on this foundation, the Guide concludes by illustrating how organizations can use this cycle of data capture and preparation to enable machine learning and AI.

SERVERLESS: THE PATH TO IT PRODUCTIVITY

Modern serverless architectures are the culmination of a series of efforts to *shrink the surface area of responsibility* that developers and IT teams must manage. Fundamentally, the goal of serverless computing is to eliminate commodified work—managing server clusters, sharding databases, load balancing, capacity planning, ensuring availability—so IT teams can focus on what matters to the business. Serverless draws a sharp distinction between *commodified IT*—the mundane maintenance work that looks roughly the same at every company—and *differentiated work* that elevates IT to a direct provider of business value.

CHAPTER 1 RECAP

1. Companies face 3 new challenges:
   - the **volume** of data being created
   - the **diversity** of data formats and sources
   - the **speed** at which consumers and internal stakeholders now expect insights.
2. Cloud computing helps companies to meet these challenges by enabling **data management at scale and speed**—without having to worry about infrastructure.
3. Specifically, businesses can start to modernize their data strategies by focusing on **cloud storage and data warehousing as a first step** in building a foundation for machine learning and AI.

CUSTOMER CASE

AIRBUS DEFENCE AND SPACE

Machine learning can unlock new possibilities for businesses—from improving customer service to more accurately streamlining operations to creating new applications and experiences. Airbus Defence and Space tested the use of Google Cloud machine learning to automate the process of detecting and correcting satellite images that contain imperfections like the presence of cloud formations.

The ability to detect patterns in satellite images—such as the difference between snow and clouds—is critical to Airbus Defence and Space's users, who depend on highly precise, up-to-date, and reliable information. Historically, this process was time-consuming, error-prone, and difficult to scale. However, using Google Cloud machine learning, the company was able to "solve a problem that has existed for decades," says Mathias Ortner, Data Analysis and Image Processing Lead.

"All satellite data collected each day are automatically processed and made readily accessible in a global imagery library that is stored in Google Cloud Platform," says Bernhard Brenner, Head of the Intelligence Business Cluster at Airbus Defence and Space. "Google Cloud Platform's global scale, low latency, and infrastructure capacities in Europe give us the required performance, flexibility, and scalability for current and future data volumes, ensuring a high level of service for our customers."[5]

READ MORE

COMPANY

Airbus Defence and Space

INDUSTRY

Aerospace

ABOUT

Airbus Defence and Space, a division of Airbus Group, is Europe's number one defense and space enterprise and the second largest space business worldwide. Its activities include space, military aircraft, and related systems and services. It employs more than 38,000 people worldwide.

CHAPTER 2

CLOUD STORAGE & DATA WAREHOUSING

02

02

CLOUD STORAGE & DATA WAREHOUSING

Centralizing raw data from key business processes into cloud storage is one of the first steps organizations can take to modernize. In doing so, they position themselves to tap analytics capabilities in the cloud.

Data silos scattered across the enterprise continue to vex business and IT teams alike, with new silos (whether for organizational or technical reasons, or both) created daily.[6] *Harvard Business Review* has published about the need for a single source of truth for data, as well as distinct lenses through which different lines of business can view the data.[7]

Cloud storage and data warehousing enable companies to do both—maintain a single, central repository, while enabling different business functions to analyze data in ways that meet their unique needs—with greater speed and flexibility than previously possible. Together, these capabilities help create a 360-degree view of the business across silos.

Capture Raw Data for Future Analysis

IDC estimates that less than 1% of all files gets analyzed.[8] The other 99%—depending on the timing of business needs—contains insights material to decision-making. Since organizations cannot predict the business questions that will arise, they need frictionless ways to store large volumes of data cheaply and flexibly. This is especially true for unstructured files, which make up the majority of data generated.[9]

With cloud, businesses can store enormous volumes of files at low cost (below one penny per gigabyte at time of writing).[10] Data that's currently needed can be kept "warm"—available globally to serve applications or to run analytics—while data with still-untapped value remains in cheaper cold storage. The most compelling online storage allows even cold archival data to be retrieved quickly with extremely low latency.

Besides saving money, cloud storage serves as the basis for powerful analytics. Businesses can capture structured and unstructured files seamlessly in their native formats. Because storage is intentionally separated from processing and analysis, teams can defer structuring raw data for analytics until business questions arise. Crucially, raw data from the same foundation can be restructured easily to answer new questions on the fly. What sets cloud storage apart is how efficiently these data-capture and repurposing steps can happen. To position an organization to benefit from analytics, teams need to ensure that raw data from their business processes is captured and centralized.

This flexibility is accelerating adoption of the cloud as a repository for organizations' unstructured data: around half of organizations across the U.S., Europe, and Asia-Pacific anticipate jumps of at least 5% in their storage of unstructured data in the cloud over the next year, with many reporting a greater than 10% increase.[11]

THE INTERNET OF THINGS

According to a survey of more than 500 global IT leaders conducted by *MIT Sloan Management Review* on behalf of Google Cloud, cloud adoption continues to accelerate, with a majority (65%) of applications, data, and/or infrastructure expected to be cloud-based by 2019.

The Internet of Things (IoT) is an important driver of this move to the cloud, with 91% of respondents with IoT initiatives either currently deploying (59%) or planning to deploy (32%) data from IoT-connected devices in the cloud. Respondents cited the ability to integrate with new tools and platforms (33%), faster app deployment and iteration (31%), increased flexibility in business processes and vendor choices (29%), and increased security (28%) as the top reasons to deploy IoT data in the cloud.

To make meaningful use of IoT data, companies must be able to understand it in context. A cloud data warehouse that allows for both batch and streaming inputs, paired with a powerful analytics platform, helps ensure that your IoT data can deliver real-time insights.

Manage Data Across Silos

Armed with the ability to capture data of any kind economically, organizations can turn their attention to enabling a disciplined view of their most important business processes. While cloud storage centralizes data in its raw native format, a cloud data warehouse enables businesses to pull together data from disparate silos for analytics—just as a traditional data warehouse would. With cloud, companies can manage large volumes of data with minimal capital investment, scale practically indefinitely, and pay only for what they use. Managed cloud services take it a step further, freeing IT from worrying about any of the underlying infrastructure. Companies must consider which business questions need answering, and the data required to answer them.

For example:

- What are the primary business goals for my data? To understand how users interact with my systems, identify trends, increase sales, build consumer loyalty, or something else?
- Where will my most important data come from (transactions, server logs, cloud services, devices/IoT, social media)? Are these imported into cloud storage already?
- How fast must my system incorporate new data in reports and visualizations?
- Is there a culture that encourages data-driven decision-making across the organization (not just among IT analysts and data scientists)? Who should have access to the analytics platform?

Once business goals are determined, companies must identify sources of input data across silos to import into a cloud data warehouse for analysis. Here's a list of typical input sources:

Cloud storage

Data from cloud storage can be imported into a cloud data warehouse for analytics.[12] At this stage, a schema can be formalized based on the business questions that need answering—bringing structure to raw data for analysis.

Analytic and transactional databases

Data stored in analytic and transactional databases can be batch-loaded or streamed row by row directly into a cloud data warehouse.

Data stored within cloud services

Data stored with popular SaaS providers can be imported into a cloud data warehouse—in many cases automatically.

Streaming data

Data from web, mobile, and IoT applications can bypass cloud storage and be streamed directly into a cloud data warehouse (see Chapter 3: Real-Time Data Integration).

Data Governance

Exponential growth in the global volume of data is not the only obstacle businesses face. According to Forrester, fast-changing requirements around analytics and reporting, as well as misalignment between business and IT, are among the top challenges impeding organizations' business intelligence efforts.[13] In addition, the well-documented data science talent gap (see "Rise of the Citizen Data Scientist," page 14) requires businesses to consider new approaches to developing analytical expertise.

With role-based access, any individual or application developer can query data stored in a cloud data warehouse, generate reports, or access visualizations. Cloud data warehousing supports individualized, need-to-know access management. Tailored access controls and complete auditability help democratize data science, while still maintaining security safeguards. Indeed, over one-half of firms across the U.S., Europe, and Asia-Pacific report they either are implementing, have implemented, or are expanding their use of self-service business intelligence tools across the enterprise.[14]

RISE OF THE CITIZEN DATA SCIENTIST

The responsibility for drawing statistically accurate conclusions based on data was once the exclusive purview of professional data scientists. But by 2018, according to McKinsey, "the U.S. alone could face a shortage of 140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts with the know-how to use the analysis of Big Data to make effective decisions."[15]

As competition intensifies, most companies will need a diversified talent strategy. *Citizen data scientists*—who, as defined in InformationWeek, are people who leverage data analytics, but whose main job functions aren't statistics or analytics—can be a powerful complement to in-house data scientists, especially for companies that invest in building a culture of data science.[16]

To be successful, would-be citizen data scientists need:

- Access to data
- Curiosity
- Facility with SQL
- Domain expertise
- Collaboration

CHAPTER 2 RECAP

1. Cloud storage allows organizations to capture both structured and unstructured data of any kind in its native format. **Centralizing data into cloud storage** creates a foundation for analytics, with details deferred until organizations have concrete business questions to ask of their data.

2. A cloud data warehouse enables organizations to **pull together data from disparate silos for analytics**, including from cloud storage, analytic and transactional databases on-prem or in the cloud, or data stored with other cloud services. Organizations can run queries, generate reports, and create visualizations—without managing the underlying infrastructure.

3. **Role-based access democratizes analytics** across an organization. A cloud data warehouse can be scoped enterprise-wide, or organized flexibly based on the structure of the organization.

CUSTOMER CASE

COLORADO CENTER FOR PERSONALIZED MEDICINE

The Colorado Center for Personalized Medicine (CCPM) is conducting breakthrough research through the analysis of patient DNA to predict disease risk and develop targeted treatments based on an individual's genetics. CCPM relies on *Health Data Compass*, CCPM's enterprise health data warehouse. Health Data Compass integrates patient genomic data from CCPM and electronic health records from UCHealth, Children's Hospital Colorado, and CU Medicine, including external records such as insurance claims, public health records, and environmental data.

Health Data Compass previously used a traditional on-premises system to store and analyze data. But that approach proved costly to maintain and didn't scale for the center's existing analytics needs, let alone its projected growth. Following a comprehensive six-month pilot project, Health Data Compass migrated to GCP and Tableau, which together can handle massive data sets and powerful visual data analyses, while costing less and allowing for easy scalability as CCPM grows. Significant to CCPM's decision was the ability of GCP, including Google Cloud's data warehouse BigQuery, to support HIPAA compliance per CCPM's requirements.

"We take our responsibility to protect patient data very seriously. Google Cloud Platform provides significant advantages in data security over on-premises systems and helps us achieve HIPAA compliance," says Michael Ames, Associate Director for Health Data Compass and Director of Enterprise Architecture for CCPM.[17]

READ MORE

COMPANY

Colorado Center for Personalized Medicine

INDUSTRY

Healthcare

ABOUT

The Colorado Center for Personalized Medicine (CCPM) is a partnership among the University of Colorado Denver, UCHealth, Children's Hospital Colorado, and CU Medicine, and is located in the Denver, Colorado, area.

CHAPTER 3

REAL-TIME DATA INTEGRATION

03

REAL-TIME DATA INTEGRATION

03

Data scientists report spending upwards of 50–80% of their time mired in the "data wrangling, data munging, and data janitor work" required to prepare data for analysis.[18] The need to provision resources and scale server clusters up and down against unpredictable workloads continues to plague teams doing data preparation on-prem.[19]

Less "Data Janitor" Work with Managed Services

Fully managed cloud services help insulate IT from the infrastructure work involved in doing large-scale data prep and data integration. Consider a smart thermostat seeking to learn and adjust to the preferences of different teams in an office building. While the thermostat is in use, the cloud ingests raw usage data, such as temperature settings and energy consumption levels throughout the day. As data comes in, a processing pipeline can be spun up on demand to prepare the raw data: ensuring inputs fall within a valid range, converting temperature and energy use into the desired units, formatting time data. The data pipeline formally structures this data, then loads the transformed results into a cloud data warehouse. Queries, visualizations, and reports are available instantly.

of companies

INDICATE INTEREST IN DEPLOYING SELF-SERVICE DATA PREPARATION TO SUPPORT BIG DATA INITIATIVES.[20]

With fully managed cloud services, the infrastructure resources required to support this workflow are allocated automatically, then spun back down afterward. Companies pay only for the resources they use, helping eliminate waste and guesswork in forecasting.

Toward Real-Time Data Analytics

While traditional systems focused on analyzing data offline "in batch," the demand for real-time insights calls for a new approach. Cloud-based streaming analytics systems are built to handle data streaming in from web applications, smartphones, or millions of IoT sensors in real time. Hundreds of thousands of sensors can be installed on field equipment to report their raw status to the cloud continuously for processing and monitoring. Visual feeds can be parsed in real time for applications like anomaly detection and facial/object recognition. With widely tested and deployed cloud services being tapped for use cases like these, streaming data analytics can be implemented in a matter of days.

With *real-time streaming data analytics*, data streams directly into processing pipelines. The transformed data can then be integrated into a cloud data warehouse—allowing for queries, visualization, and reporting within seconds. In this way, the processing pipeline serves as a kind of middleware that can be spun up on demand, able to join data streaming in real time with batch data pulled in from storage. Data can be structured flexibly to answer an organization's business questions as they arise.

Organizations therefore have two complementary paths—batch and streaming—by which they can capture, prepare, and integrate data from any source to any target. Managed cloud services make it possible to seamlessly accomplish both.

MAKING THE MOST OF YOUR EXISTING BIG DATA INVESTMENTS

Many forward-looking enterprises are already using Big Data, often based on open-source tools like Apache Hadoop and Apache Spark. For these businesses, it's possible to protect existing investments in talent and tools—while still achieving the cloud's productivity advantages.

Adoption of open-source Big Data tools is widespread—and growing. Globally, many firms are storing an increasing amount of unstructured data in public cloud file systems (including Hadoop). Over one-third of respondents in the U.S. and Europe—and well over half in Asia-Pacific—report that they're implementing, have implemented, or are expanding their implementation of Hadoop (including HBASE, Accumulo, MapR, Cloudera, Hortonworks). Similarly, around one-third of respondents in the U.S. and Europe (and a whopping 60% in Asia-Pacific) are implementing, have implemented, or are expanding their implementation of in-memory data platforms (including Apache Spark, SAP Hana, Kognitio, Terracotta, Gigaspaces).

For organizations like these, the cloud offers two primary options:

- Continue to manage Big Data projects using familiar open-source tools—but migrate to virtual machines in the cloud. The usual cloud benefits apply: retire expensive CapEx; move to an OpEx model of billing, where organizations pay according to data stored and processed; scale seamlessly. Note that in this forklift model, developers and IT teams are still required to manage their own storage and data processing pipelines. However, it is the most straightforward route for leveraging talent, tools, and vendor relationships already in place.
- The cloud offers fully managed versions of many of the most popular open-source tools in Big Data. For example, running Apache Hadoop, Apache Spark, Apache Pig, and Apache Hive in the cloud offloads basic data-management tasks like deployment, logging, and monitoring.[21] This is an excellent option for teams looking for the best of both on-prem and cloud-native worlds.

Either of these options lets organizations protect their existing investments in deploying Big Data—but smartly use cloud economics to control costs and gain flexibility.

CHAPTER 3 RECAP

1 **Cloud-based data processing pipelines** enable organizations to extract, transform/prepare, and integrate data from any source to any target (on-prem or cloud).

2 **Serverless approaches to data preparation** fully manage the underlying infrastructure; resources are allocated automatically based on the needs of each data processing pipeline.

3 **Cloud streaming analytics** allow data from web, mobile, and IoT applications to stream into data processing pipelines in real time. From here, data can be prepared and integrated into a cloud data warehouse to support a real-time view of the business.

CASE STUDY

CITIBANK UK

In this proof of concept, the team's task was to show how easy it would be for Citibank to use Google BigQuery and Google Cloud Pub/Sub to analyze and consume roughly 1,000 financial instruments' worth of historic and near-real-time tick data from Thomson Reuters. The work was done in collaboration with Sean Micklethwaite, lead developer from Citibank, and Sebastian Fuchs, solution specialist from Thomson Reuters.

"We wanted an API that we could query for historic data on demand, without the need to maintain our own data warehouse, and all the cost and operational overhead that entails," Micklethwaite explains. "Additionally, we required real-time updates to market data with human-level latency. With Google Cloud, we get access to all the data we need through a single platform. BigQuery takes care of our historic tick data needs, and has the power to process raw ticks at high frequency over large ranges. Cloud Pub/Sub takes care of our real-time data requirements, and we receive all data in a consistent format."

Adds Fuchs: "We started to use BigQuery without having the need to do excessive capacity planning up front. It simply grows as needed, both from a content provisioning [perspective] as well as from a number-of-user-queries point of view."

READ MORE

COMPANY

Citibank UK

INDUSTRY

Financial Services

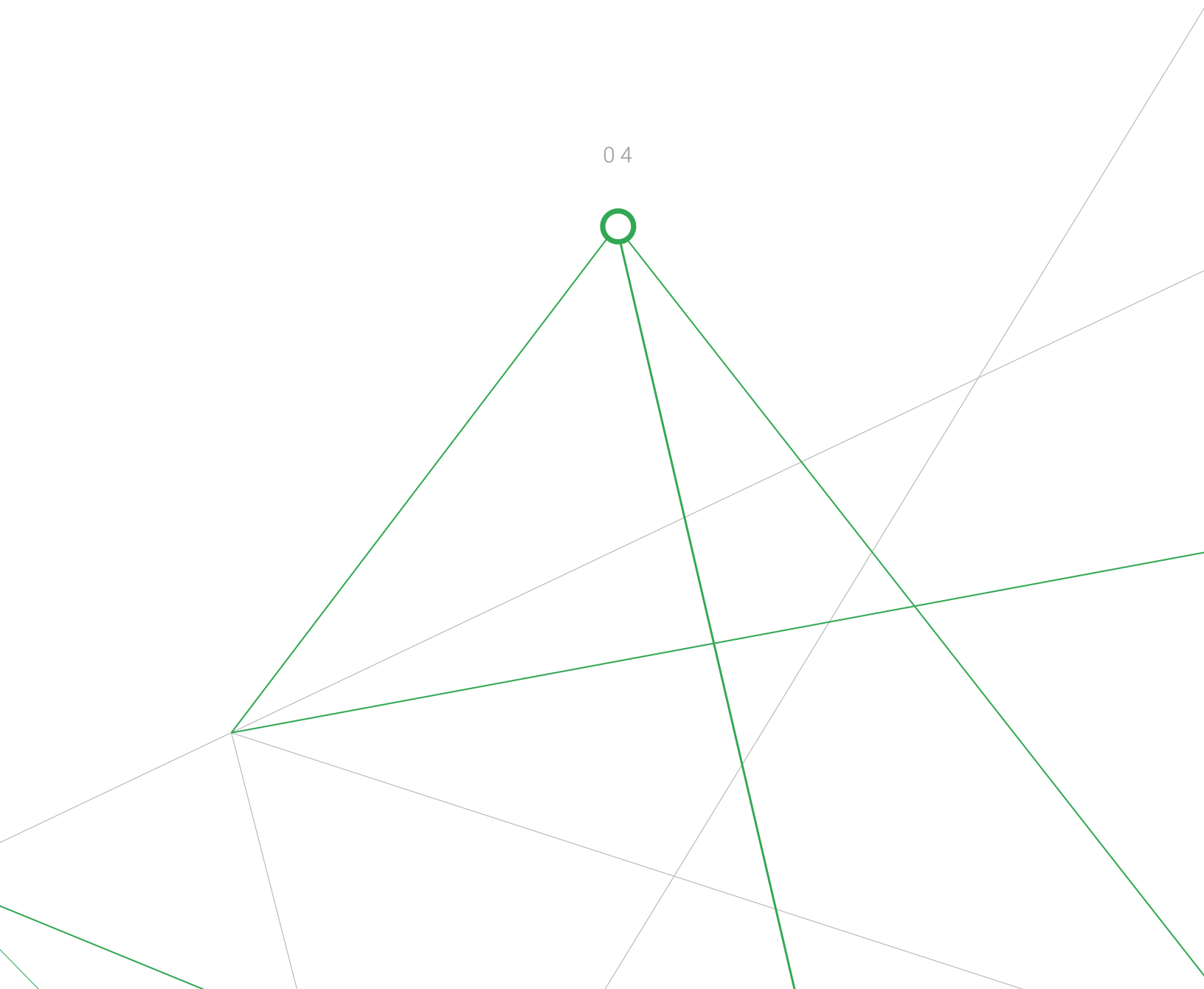
ABOUT

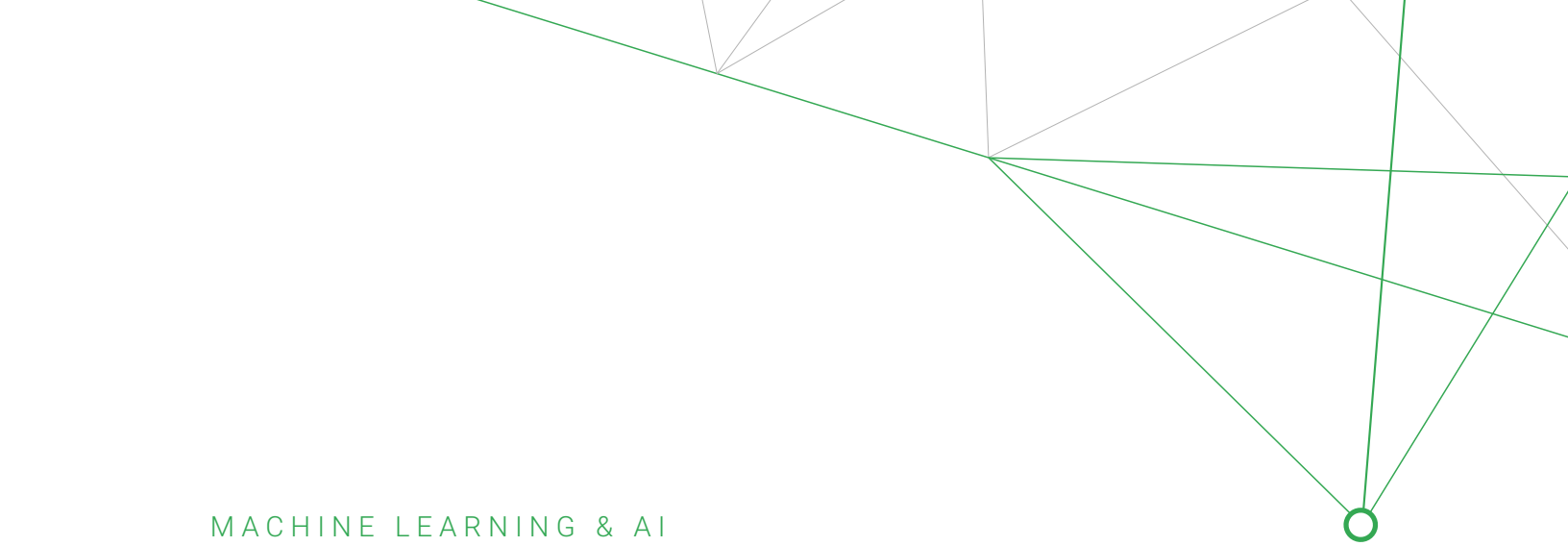
In a proof-of-concept experiment, Google Cloud partnered with Thomson Reuters to demonstrate to Citibank's Global Banking and Markets department the benefits of combining Google's core data technologies with Thomson Reuters financial market content.

CHAPTER 4

MACHINE LEARNING & AI

04





04

MACHINE LEARNING & AI

Recent breakthroughs in machine learning (ML) and artificial intelligence (AI) frequently make headlines. Computers have bested human world champions in Go, a board game with more positions than there are atoms in the universe.[22] They've mastered popular video games and, critically, learned to recognize cats.[23] More recently, an AI effort achieved massive savings in energy costs, highlighting machine learning as "a general-purpose framework to understand complex dynamics."[24] This framework is starting to find diverse applications—and deliver results—across many industries.

The concept of AI is simple: the ability for software to improve without needing to be explicitly programmed. Rather than requiring developers to write new code manually, AI relies on algorithms capable of getting "smarter" by ingesting more real-world data. Centralizing the storage and preparation of data in the cloud—the goals of Chapters 2 and 3, respectively—creates the ideal foundation for training and improving AI models.

The AI opportunity extends beyond simply automating once-manual tasks. In online retail, for example, machine-learning algorithms can ingest and analyze enormous volumes of consumer data as potential buyers navigate through a retailer's online store or mobile app. The more data the model ingests, the closer it comes to understanding exactly when—and why—a particular buyer will decide to make a particular purchase. Eventually, this learning becomes predictive, enabling the retailer to surface the right product for the right person at the right time. This level of personalization—once typified by the small-town shopkeeper who knew the names and birthdays of her customers' children—is now possible at scale.

of respondents BELIEVE THEIR ORGANIZATION'S FUTURE SUCCESS DEPENDS ON THE SUCCESSFUL IMPLEMENTATION OF MACHINE LEARNING.[25]

Many small, tech-forward companies are already seeing results from ML—but more established enterprises have a unique opportunity to tap into a wealth of historical data.[26] With ML, outcomes depend on the sheer quantity of data available to feed into training models (see "Pre-Trained Models: A First Step Toward AI"). Established companies can tap their first-party data—everything from IT system logs to financial transactions to customer-service call transcripts—to train and optimize these models, returning insights unique to the company.

PRE-TRAINED MODELS: A FIRST STEP TOWARD AI

The most straightforward way to get started with AI is to use pre-trained machine learning models, available instantly through the cloud. No prior knowledge of ML is required. These capabilities may be familiar to those who use popular consumer applications, where some of the models have reached levels of predictive accuracy that exceed human ability:

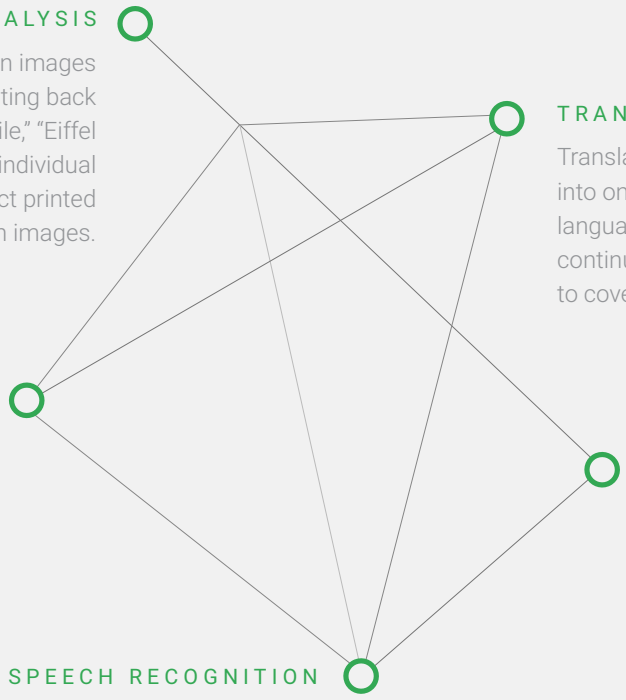


IMAGE ANALYSIS

Understand the content within images by supplying images and getting back text categories (e.g., "automobile," "Eiffel Tower"). Intelligently detect individual objects and faces, or even extract printed words contained within images.

TRANSLATION

Translate any provided string into one of many supported languages. Translation services continuously improve and expand to cover new languages.

VIDEO ANALYSIS

Identify objects and meaningful entities within videos; understand what's going on at the level of an individual shot or frame.

TEXT ANALYSIS

Understand the structure and semantic meaning of text. Extract information about people, places, and events from any text. Understand sentiment and parse information from news articles, blogs, customer support conversations/chat logs, or any text you supply.

SPEECH RECOGNITION

Convert audio to text with extraordinary accuracy. Supply audio to the API and get full text returned instantly, with support for many languages. Enable real-time use cases like transcription, or create applications driven by voice.

These services are general (not tied to consumer applications), and can be integrated into any application easily via simple API calls. Developers don't need to know any of the underlying details. Without having to develop any of these services in-house, companies can tap the latest capabilities instantly, as a service.

Established enterprises and industry incumbents typically have decades of first-party data accumulated: financial transactions; system logs; raw data generated from manufacturing, retail, and e-commerce data captured over years; performance results from marketing campaigns. Properly refined and used to train custom machine learning models, this data becomes a source of predictive power. Rather than repurposing canned services, established companies can use first-party data to optimize their business processes for their customers—a potent source of differentiation.

Use cases span many industries and reveal some of AI's most promising applications. Fraud detection in financial services and preventive maintenance in manufacturing highlight the ability to surface anomalies from a sea of transactions and messy logs—a common need in many domains. Diagnosis and treatment suggestions in healthcare and judgments on creditworthiness highlight machine learning's ability to assist with categorization—also generally useful.

QUANTIFYING THE PAYOFF

In partnership with the research firm M-Brain, Google Cloud surveyed 20 IT and business leaders who have implemented machine learning projects about key benefits yielded from the projects. Top benefits named included:

- Time savings
- Cost savings
- Better risk management
- Improved quality of analytics
- Increased revenues

Others listed automation, improved service, and improved inventory planning.[27]

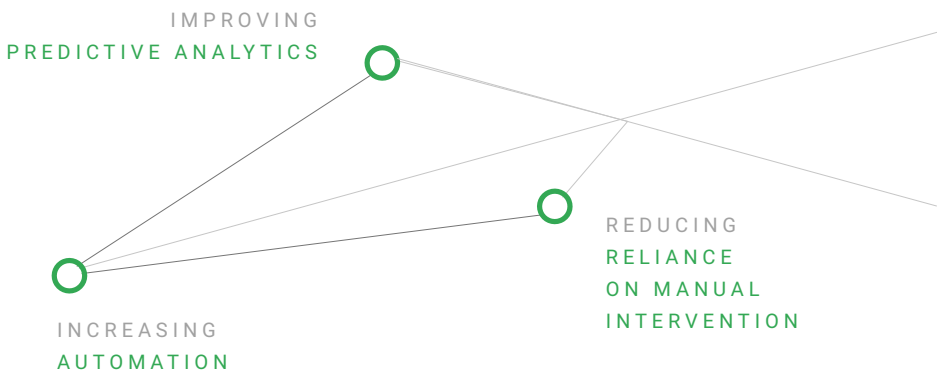
Virtuous Cycle: Capture, Prepare, Train, Predict

The capabilities introduced in Chapters 2 and 3 serve as a foundation for training machine learning models using first-party data. With raw data already centralized in cloud storage and a cloud data warehouse, serverless data pipelines can continuously extract this data and prepare it to train bespoke ML models. Since ML models can themselves be housed in the cloud, they are immediately available to applications to make predictions. This loop forms a virtuous cycle, in which ML models housed in the cloud keep improving from new training data, which in turn keeps the models fresh and relevant.

TOP REPORTED BUSINESS NEEDS

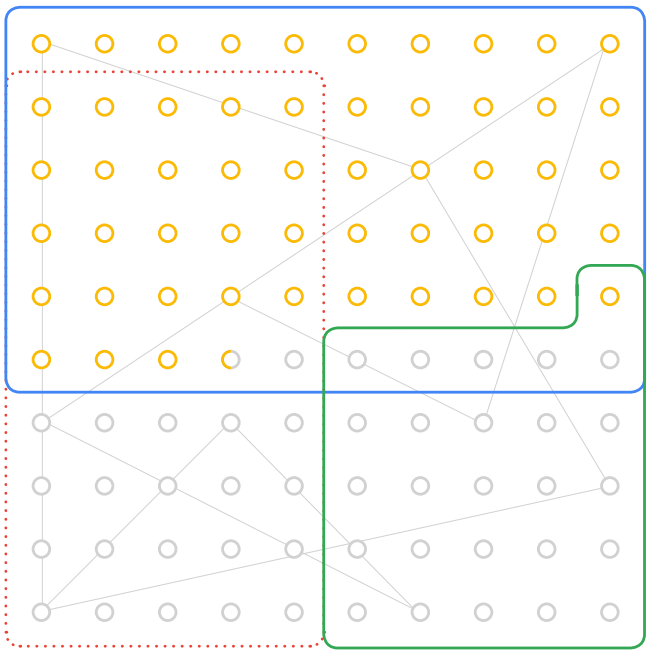
across:[28]

- HEALTHCARE
- FINANCIAL SERVICES
- MANUFACTURING
- RETAIL
- MEDIA/GAMING



ML: THE NEW PROVING GROUND FOR COMPETITIVE ADVANTAGE

The age of machine learning has finally arrived—and it's already in full swing within smaller, tech-forward companies, according to a new survey of business and technology leaders by *MIT Technology Review Custom*. Some key findings:[29]



60% of respondents have already **implemented machine learning strategies.**

>50% of **early-stage ML implementers are already seeing ROI.**

45% have **achieved more extensive data analysis and insights.**

26% report a **stronger competitive advantage.**

Download the complete report here.

CHAPTER 4 RECAP

1. **AI, and its subset of machine learning, are simple in concept:** the ability for software to improve without needing to be explicitly programmed.
2. **AI relies on large volumes of training data**, giving established businesses a unique advantage to pull from the wealth of business data generated over long operating histories.
3. **Cloud storage, data warehousing, data integration, and analytics provide a natural foundation** for AI and ML by making data available for real-time training and optimization, powering a virtuous cycle of continuous improvement.

CONCLUSION

In an age of abundant data and immediate answers, the ability to extract value from data—regardless of source, size, and requirements around timeliness—will be at the heart of an organization's competitive advantage.

The first step is to rethink data strategy from the ground up. Today's cloud tools allow companies to manage enormous volumes of diverse data types more efficiently, at lower cost, than previously possible. Businesses that take a modern approach to capturing, storing, preparing, and analyzing their data will have the foundation to take advantage of machine learning and AI. Ultimately, these new capabilities will translate into closer relationships between companies and their customers, enabling businesses to be more predictive in every interaction.

LEARN MORE ABOUT WHAT GOOGLE CLOUD CAN DO FOR YOUR BUSINESS.

Storage & Databases

LEARN MORE

Big Data Solutions

LEARN MORE

Machine Learning & AI

LEARN MORE

WORKS CITED

1. Eighty-one percent of senior executives surveyed by Ernst & Young agreed that data should be at the heart of decision-making, only 31% had significantly restructured operations to incorporate big data, and a mere 23% had implemented organization-wide data strategies. Ernst & Young, *Becoming an Analytics-Driven Organization* (2015) (link).
2. David Reinsel et al., *Data Age 2025: The Evolution of Data to Life-Critical* (IDC, 2017) (link).
3. Cade Metz, "Exclusive: Inside Google Spanner, the Largest Single Database on Earth," *Wired* (26 November 2012) (link).

   Cade Metz, "Spanner, the Google Database that Mastered Time, Is Now Open to Everyone," *Wired* (14 February 2017) (link).
4. Robert McMillan, "Inside the Artificial Brain that's Remaking the Google Empire," *Wired* (16 July 2014) (link). TensorFlow (link).
5. Airbus Defence and Space, "Airbus Defence and Space Selects Google Cloud Platform as Preferred Partner," news release (18 October 2016) (link).
6. Forrester, *Forrester's Global Business Technographics Data and Analytics Survey* (2016) (link).
7. Leandro DalleMule and Thomas H. Davenport, "What's Your Data Strategy?" *Harvard Business Review* (May 2017) (link).
8. John Gantz and David Reinsel, *The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East* (IDC, 2012) (link).
9. Tracie Kambies et al., *Tech Trends 2017: Dark Analytics: Illuminating Opportunities Hidden within Unstructured Data* (Deloitte University Press, 2017) (link).
10. *Google Cloud Storage Pricing*, Google Cloud Platform (link).
11. Forrester, *Forrester's Global Business Technographics Data and Analytics Survey* (2016) (link).
12. Modern cloud data warehouses support importing (and even ad-hoc querying) many semistructured formats automatically. For unstructured data that needs to be transformed first (e.g., ETL), see Chapter 3: Real-Time Data Integration.
13. Forrester, *Forrester's Global Business Technographics Data and Analytics Survey* (2016) (link).
14. Forrester, *Forrester's Global Business Technographics Data and Analytics Survey* (2016) (link).
15. James Manyika et al., *Big Data: The Next Frontier for Innovation, Competition, and Productivity* (McKinsey Global Institute, 2011) (link).
16. Lisa Morgan, "Citizen Data Scientists: 7 Ways to Harness Talent," *InformationWeek* (24 July 2015) (link).
17. *Colorado Center for Personalized Medicine: Improving Healthcare by Integrating Patient Records and Genetic Data Using Google Cloud Platform and Tableau* (Google Cloud Platform, 2017) (link).
18. Steve Lohr, "For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights," *The New York Times* (17 August 2014) (link).
19. Forrester, *Forrester's Global Business Technographics Data and Analytics Survey* (2016) (link).
20. Forrester, *Forrester's Global Business Technographics Data and Analytics Survey* (2016) (link).
21. *Apache Hadoop*, The Apache Software Foundation (link).

    Apache Spark, The Apache Software Foundation (link).

    Apache Pig, The Apache Software Foundation (link).

    Apache Hive, The Apache Software Foundation (link).

22. Paul Mozur, "Google's A.I. Program Rattles Chinese Go-Master As It Wins Match," *The New York Times* (25 May 2017) (link).

23. Nicola Twilley, "Artificial Intelligence Goes to the Arcade," *The New Yorker* (25 February 2015) (link).

    John Markoff, "How Many Computers to Identify A Cat? 16,000," *The New York Times* (25 June 2012) (link).

24. James Vincent, "Google Uses DeepMind AI to Cut Data Center Energy Bills," *The Verge* (21 July 2016) (link).

25. *Harvard Business Review Analytic Services Global Data and Analytics Survey*, sponsored by Google (2017).

26. A survey by *MIT Technology Review* shows smaller organizations well on their way to machine learning adoption and returns: 60% of a pool of 375 respondents in which nearly two-thirds were companies with fewer than 1,000 employees drawn largely from the technology, business, and financial services industries. *MIT Technology Review Custom* and Google Cloud, *Machine Learning: The New Proving Ground for Competitive Advantage* (2017) (link).

27. Anna Rader, *Machine Learning Initiatives Across Industries: Practical Lessons from IT Executives* (M-Brain, sponsored by Google, 2017) (link).

28. Anna Rader and Irida Jano, *Machine Learning Market Research: How Leading Industries Are Adopting AI* (M-Brain, 2017) (link).

29. *MIT Technology Review Custom* and Google Cloud, *Machine Learning: The New Proving Ground for Competitive Advantage* (2017) (link).

1600 Amphitheatre Parkway, Mountain View, CA 94043