too BIG to IGNORE

THE BUSINESS CASE FOR BIG DATA

PHIL SIMON Award-Winning Author of THE AGE OF THE PLATFORM

WILEY

Contents

CHAPTER 1

Data 101 and the Data Deluge

Any enterprise CEO really ought to be able to ask a question that involves connecting data across the organization, be able to run a company effectively, and especially to be able to respond to unexpected events. Most organizations are missing this ability to connect all the data together.

—Tim Berners Lee

Today, data surrounds us at all times. We are living in what some have called *the Data Deluge*.[1] *Everything* is data. There's even data about data, hence the term *metadata*. And data is anything but static; it's becoming bigger and more dynamic all the time. The notion of data is somewhat different and much more nuanced today than it was a decade ago, and it's certainly much larger.

Powerful statements like these might give many readers pause, scare some others, and conjure up images of *The Matrix*. That's understandable, but the sooner that executives and industry leaders realize this, the quicker they'll be able to harness the power of Big Data and see its benefits. As a starting point, we must explore the very concept of data in greater depth—and a little history is in order. If we want to understand where we are now and where we are going, we have to know how we got here.

This chapter discusses the evolution of data in the enterprise. It provides an overview of the types of data that organizations have at their disposal today. It answers questions like these: How did we arrive at the Big Data world? What does this new world look like? We have to answer questions like these before we can move up the food chain. Ultimately, we'll get to the big question: how can Big Data enable superior decision-making?

THE BEGINNINGS: STRUCTURED DATA

Make no mistake: corporate data existed well before anyone ever turned on a proper computer. The notion of data didn't even arrive years later, when primitive accounting systems became commercially viable. So why weren't as many people talking about data thirty years ago? Simple: because very little of it was easily (read: electronically) available.

Before computers became standard fixtures in offices, many companies paid employees via manual checks; bookkeepers manually kept accounting ledgers. The need for public companies to report their earnings on quarterly and annual bases did not start with the modern computer. Of course, thirty years ago, organizations struggled with this type of reporting because they lacked the automated systems that we take for granted today. While calculators helped, the actual precursor to proper enterprise systems was VisiCalc. Dan Bricklin invented the first spreadsheet program in the mid-1970s, and Bob Frankston subsequently refined it.

In the mid-1980s, user-facing or front-end applications like manufacturing resource planning (MRP) and enterprise resource planning (ERP) systems began to make inroads. At a high level, these systems had one goal: to automate standard business processes. To achieve this

Table 1.1 Simple Example of Structured Customer Master Data

CustomerID	CustomerName	ZipCode	ContactName
1001	Bally's	89109	Jon Anderson
1002	Bellagio	89109	Geddy Lee
1003	Wynn Casino	89109	Mike Mangini
1004	Borgata	08401	Steve Hogarth
1005	Caesar's Palace	89109	Brian Morgan

goal, enormous mainframe databases supported these systems. For the most part, these systems could only process *structured data* (i.e., "orderly" information relating to customers, employees, products, vendors, and the like). A simple example of this type of data is presented in Table 1.1.

Now a master customer table can only get so big. After all, even Amazon.com "only" serves 300 or 400 million customers—although its current internal systems can support many more times that number. Tables get much longer (not wider) when they contain *transactional* data like employee paychecks, journal entries, or sales. For instance, consider Table 1.2.

In Table 1.2, we see that many customers make multiple purchases from a company. For instance, I am an Amazon customer, and I buy at least one book, DVD, or CD per week. I have no doubt that each sale represents an individual record in a very long Amazon database table somewhere. (Amazon uses this data for two reasons: [1] process my payments; and [2] learn more about my purchasing habits and recommend products that, more often than not, I consider buying.)

Table 1.2 Simple Example of Transactional Sales Data

OrderNbr	CustomerID	ProductID	OrderDate	ShipDate
119988	1001	2112	1/3/13	1/6/13
119989	1002	1234	1/6/13	1/11/13
119990	1001	2112	1/6/13	1/9/13
119991	1004	778	1/6/13	1/12/13
119992	1004	999	1/7/13	1/15/13

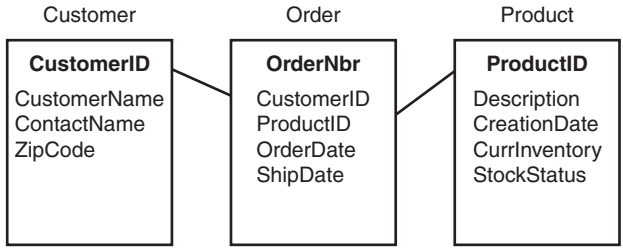


Figure 1.1 Entity Relationship Diagram (ERD)

Things are orderly under a *relational data model*. All data is stored in proper tables, and each table is typically joined with at least one other. Each table is its own entity. An *Entity Relationship Diagram* (ERD) visually represents the relationships between and among tables. A simple example of an ERD is shown in Figure 1.1.

Note that the ERD in Figure 1.1 is nothing like what you'd find behind the scenes in most large organizations. It's common for enterprise systems to contain *thousands* of individual tables (including some customized ones), although not every table in a commercial off the shelf (COTS) system contains data. Also, querying data from multiple tables requires JOIN statements. While you can theoretically query as many data sources and tables as you like (as long as they are properly joined), queries with a high number of huge tables tend to take a great deal of time to complete.* Queries improperly or inefficiently written can wreak havoc across an entire enterprise.

Throughout the 1990s and early 2000s, more and more organizations deployed systems built upon this relational data model. They uprooted their legacy mainframe systems and supplanted them with contemporary enterprise applications. Importantly, these applications were powered by orderly and expensive relational databases like Oracle, SQL Server, and others. What's more, organizations typically converted their legacy data to these new systems by following a process called *ETL* (extract, transform, and load).**

* Trust me. I've written tens of thousands of queries in my day.

** We'll see in Chapter 4 that ETL isn't really beneficial in a world of Hadoop and NoSQL because much data is far less structured these days.

Like their predecessors, ERP and CRM systems excelled at handling structured data, performing essential business functions like paying vendors and employees, and providing standard reports. With these systems, employees could enter, edit, and retrieve essential enterprise information. Corporate intranets, wikis, and knowledge bases represented early attempts to capture unstructured data, but most of this data was internal (read: generated by employees, not external entities). For the most part, intranets have not displaced e-mail as the *de facto* killer app inside many large corporations.

When asked about data, most people still only think of the structured kind mentioned in this section. "The relational model has dominated the data management industry since the 1980s," writes blogger Jim Harris on the Data Roundtable. That model "foster(s) the long-held belief that data has to be structured before it can be used, and that data should be managed following ACID (atomicity, consistency, isolation, durability) principles, structured primarily as tables and accessed using structured query language (SQL)."[2] Harris is spot-on. The relational data model is still very important, but it is no longer the only game in town. It all depends on the type and source of data in question.

Even in a Big Data world, transactional and structured data and the relational databases behind them are far from irrelevant. But organizations need to start leveraging new data sources and solutions.

STRUCTURE THIS! WEB 2.0 AND THE ARRIVAL OF BIG DATA

While business information is as old as capitalism itself, the widespread use of corporate data is a relatively recent development. The last section demonstrated how, in the 1980s and 1990s, relational databases, ERP and CRM applications, business automation, and computers all helped popularize the contemporary notion of data. Over the past few decades, organizations have begun gradually spending more time, money, and effort managing their data, but these efforts have tended to be mostly internal in nature. That is, organizations have focused on what the data generated by their own hands.

In or around 2005, that started to change as we entered Web 2.0—aka *the social web*. As a direct result, the volume, variety, and velocity of data rose exponentially, especially consumer-driven data that is, for the most part, *external* to the enterprise. The usual suspects include the rise of nascent social networks like Classmates.com, MySpace, Friendster, and then a little Harvard-specific site named The Facebook. Photo sharing began to go mainstream through sites like Flickr, eventually gobbled up by Yahoo! Blogging started to take off—as did micro-blogging sites like Twitter a few years later. More and more people began walking around with increasingly powerful smartphones that could record videos. Enter the citizen journalist. Sites like YouTube made video sharing easy and extremely popular, prompting Google to pay $1.65 billion for the company in 2007. Collectively, these sites, services, and advancements led to proliferation of unstructured data, semi-structured data, and metadata—the majority of which was external to the enterprise.

To be sure, many organizations have seen their structured and transactional data grow in velocity and volume. As recently as 2005, *Information Management* magazine estimated the largest data warehouse in the world at 100 terabytes (TB) in size. As of September 2011, Walmart, the world's largest retailer, logged "one million customer transactions per hour and fed information into databases estimated at 2.5 petabytes in size."[3] (I'll save you from having to do the math. This is 25 times as big.) For their part, today companies like Amazon, Apple, and Google are generating, storing, and accessing much more data now than they did in 2005. This makes sense. As Facebook adds more users and features, Apple customers download more songs and apps, Google indexes more web pages, and Amazon sells more stuff, each generates more data.

However, it's essential to note that Web 2.0 did not increase internal IM demands at every organization. Consider a medium-sized regional hospital for a moment. (I've consulted at many in my career.) Let's say that, on a typical day, it receives 200 new patients. For all of its transformative power, the Internet did not cause that hospital's daily patients to quadruple. Hospitals only contain so many beds. Think of a hospital as an anti-Facebook, because it faces fairly strict limits with respect to scale. Hospitals don't benefit from network effects. (Of course, they can always expand their physical space, put in more beds, hire more

employees, and the like. These activities will increase the amount of data the hospitals generate, but let's keep it simple in this example.)

Unstructured Data

We know from Tables 1.1 and 1.2 that structured data is relational, orderly, consistent, and easily stored in spreadsheets and database tables. Unstructured data is its inverse. It's big, nonrelational, messy, text laden, and not easily represented in traditional tables. And unstructured data represents most of what we call *Big Data*. According to ClaraBridge, a leader in sentiment and text analytics software, "Unstructured information accounts for more than 80 percent of all data in organizations."[4] By some estimates, unstructured data is growing ten to fifty times faster than its structured counterpart.

While everyone agrees on the growth of data, there's plenty of disagreement on the precise terminology we should be using. Some believe that unstructured data is in fact a contradiction in terms.[5] And then there's Curt Monash, Ph.D., a leading analyst of and strategic advisor to the software industry. He defines *poly-structured data* as "data with a structure that can be exploited to provide most of the benefits of a highly structured database (e.g., a tabular/relational one) but cannot be described in the concise, consistent form such highly structured systems require."[6] Debating the technical merits of different definitions isn't terribly important for our purposes. This book uses the term *unstructured data* in lieu of *poly-structured data*. It's just simpler, and it suits our purposes just fine.

Semi-Structured Data

The rise of the Internet and the web has led not only to a proliferation of structured and unstructured data. We've also seen dramatic increases in two other types of data: semi-structured data and metadata. Let's start with the former.

As its name implies, semi-structured data contains characteristics of both its structured and unstructured counterparts. Examples include

- Extensible Markup Language (XML) and other markup languages

- E-mail
- Electronic Data Interchange (EDI), a particular set of standards for computer-to-computer exchange of information

Many people are using semi-structured data whether they realize it or not. And the same holds true for metadata, discussed next.

Metadata

> The information about the package is just as important as the package itself.
>
> **Fred Smith, Founder and CEO of FedEx, 1978**

Now let's move on to metadata, a term that is increasingly entering the business vernacular. As the quote indicates, people like Fred Smith grasped its importance thirty-five years ago. The term *metadata* means, quite simply, data about data.

We are often creating and using metadata whether we realize it or not. For instance, my favorite band is Rush, the Canadian power trio still churning out amazing music after nearly forty years. While I usually just enjoy the music at concerts, sing along, and air drum,* I occasionally take pictures. Let's say that when I get home, I upload them to Flickr, a popular photo-sharing site. (Flickr is one of myriad companies that extensively use tags and metadata. Many stock photo sites like iStockphoto and Shutterstock rely heavily upon metadata to make their content easily searchable. In fact, I can't think of a single major photo site that doesn't use tags.)

So I can view these photos online, but what if I want other Rush fans to find them? What to do? For starters, I can create an album titled *Rush 2012 Las Vegas Photos*. That's not a bad starting point, but what if someone wanted to see only recent pictures of the band's insanely talented drummer Neil Peart? After all, he's not in all of my pictures, and it seems silly to make people hunt and peck. The web has evolved, and so has search. No bother; Flickr has me covered. The site encourages me to tag pictures with as many descriptive labels as I like, even offering suggestions based upon similar photos

* To see what I mean, Google "Rush Car Commercial Fly By Night 2012."

and albums. In the end, my photos are more findable throughout the site for everyone, not just me. For instance, a user searching for "Las Vegas concerts" or "rock drummers" may well come across a photo of Peart in action in Las Vegas, whether she was initially looking for Rush or not.

In this example, the tags are examples of metadata: they serve to describe the actual data (in this case, the photo and its "contents"). But these photos contain metadata whether I choose to tag them or not. When I upload each photo to Flickr, the site captures each photo's time and date and my username. Flickr also knows the size of the photo's file (its number of KBs or MBs). This is more metadata. And it gets better. Perhaps the photo contains a date stamp and GPS information from my camera or smartphone. Let's say that I'm lazy. I upload my Rush photos in mid-2013 and tag them incorrectly as "Cleveland, Ohio." Flickr "knows" that these photos were actually taken in November 2012 in Las Vegas and kindly makes some recommendations to me to improve the accuracy of my tags and description.* In the future, maybe Flickr will add facial recognition software so I won't have to tag anyone anymore. The site will "learn" that my future Rush photos will differ from those of a country-music-loving, conservative Republican who adores radio show host Rush Limbaugh. (For more on how tagging works and why it's so important, check out *Everything Is Miscellaneous: The Power of the New Digital Disorder* by David Weinberger.)

Because of its extensive metadata, Flickr can quickly make natural associations among existing photos for its users. For instance, Flickr knows that one man's "car" is another's "automobile." It also knows that maroon and mauve are just different shades of red and purple, respectively. This knowledge allows Flickr to provide more accurate and granular search results (see Figure 1.2).

If I want to find photos taken of Neil Peart from 6/01/2012 to 10/01/2012 in HD only with the tag of "Clockwork Angels" (the band's most recent studio album), I can easily perform that search. (Whether I'll see any results is another matter; the more specific the criteria, the

* The site uses Exif data (short for Exchangeable Image File), a standard format for storing interchange information in digital photography image files using JPEG compression.

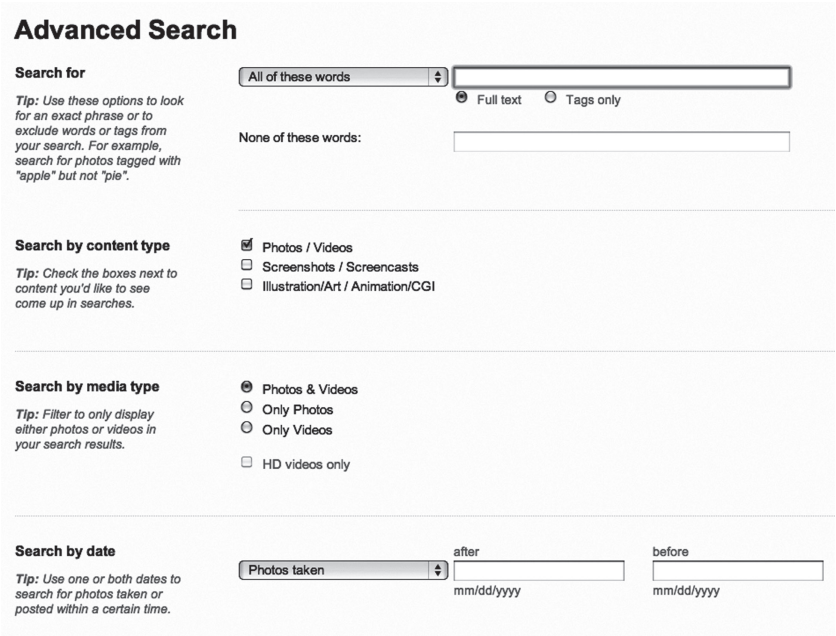


Figure 1.2 Flickr Search Options
Source: Flickr.com

less likely that I'll see any results.) The larger point is that, without metadata, searches like these just aren't possible.

Smartphones with GPS functionality make tagging location easier than ever—and just wait until augmented reality comes to your smartphone. If not the final frontier, the next logical step in tagging is facial recognition. To this end, in June 2012, Facebook acquired facial recognition start-up Face.com for an undisclosed sum (rumored to be north of $100 million). The Tel Aviv, Israel-based start-up "offers application programming interfaces (APIs) for third-party developers to incorporate Face.com's facial-recognition software into their applications. The company has released two Facebook applications: Photo Finder, which lets people find untagged pictures of themselves and their Facebook friends, and Photo Tagger, which lets people automatically bulk-tag photos on Facebook."[7]

So from a data perspective, what does all of this mean? Several things. First, even an individual photo has plenty of metadata

associated with it—and that data is stored somewhere. Think about the billions of photos online, and you start to appreciate the amount of data involved in their storage and retrieval. Second, photos today are more complex because they capture more data than ever—and this trend is only intensifying.

Let's get back to my Rush example. If I look at my concert pictures from last night, I'm sure that I'll find one with poor focus. Right now, there's not much that I can do about it; Photoshop can only do so much. But soon there might be hope for fuzzy pictures, thanks to the folks at the start-up Lytro. Through light field technology, Lytro's cameras ultimately "allow users to change the focus of a picture after the picture is taken."[8] Plenoptic cameras such as Lytro's represent "a new type of camera that dramatically changes photography for the first time since the 1800s. [It's] not too far away from those 3D moving photographs in the *Harry Potter* movies."[9]

THE COMPOSITION OF DATA: THEN AND NOW

Over the past decade, many organizations have continued to generate roughly the same amount of internal, structured, and transactional data as they did before the arrival of Web 2.0. For instance, quite a few haven't seen appreciable changes in the same number of employee and vendor checks cut. They book roughly the same number of sales and generate a more or less stable number of financial transactions. The "data world" *inside* of the organization in many cases has not changed significantly. However, this is in stark contrast to the data world outside—and *around*—the enterprise. It could not be more different. Consider Figure 1.3.

As Figure 1.3 shows, there is now much more external and unstructured data than its structured counterpart—and has been for a long time. At the same time, the amount of structured, transactional data has grown exponentially as well. The ostensible paradox can be explained quite simply: while the amount of structured data has grown fast, the amount of unstructured data has grown much faster. Analytics-as-a-service pioneer 1010data "now hosts more than 5 trillion—yes, trillion with a "t"—records for its customers."[10] Despite statistics like this, most data today is of the unstructured variety.

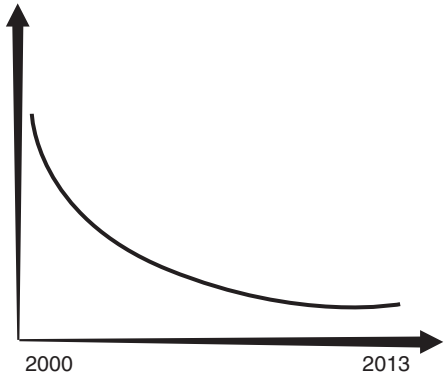


Figure 1.3 The Ratio of Structured to Unstructured Data

As mentioned before, unstructured data has always existed, even though few people thought of it as "data." So other than the terminology, what's different now? First, as mentioned earlier, there's just more unstructured data today than at any point in the past. Second, much of this unstructured data is digitized and available nearly *instantly*—especially if its owners want it to be.* Delays have evaporated in many cases, although many organizations cannot access real-time data. Be that as it may, by and large, today unstructured data doesn't need to be transcribed, scanned into computers, or read by document storage systems. *Data is born digital*—or at least it can be. (There are still plenty of hospitals and doctors' offices that refuse to embrace the digital world and electronic medical records.)

Consider books, newspapers, and magazines for a moment—all good examples of unstructured data (both 20 years ago and now). For centuries, they were released only in physical formats. With the rise of the Internet, this is no longer the case. Most magazines and newspapers (sites) are available electronically, and print media has been dying for some time now. Most proper books (including this one) are available both in traditional and electronic formats. There is no time lag. In fact, some e-books and Kindle singles are *only* available electronically.

* Even this isn't entirely true, as Julian Assange has proven.

THE CURRENT STATE OF THE DATA UNION

Unstructured data is more prevalent and bigger than ever. This does not change the fact that relatively few organizations have done very much with it. For the most part, organizations lamentably have turned a deaf ear to this kind of information—and continue to do so to this day. They have essentially ignored the gigantic amounts of unstructured or semi-structured data now generated by always-connected consumers and citizens. They treat data as a four-letter word.

It's not hard to understand this reluctance. To this day, many organizations struggle managing just their own transactional and structured data. Specific problems include the lack of master data, poor data quality and integrity, and no semblance of data governance. Far too many employees operate in a vacuum; they don't consider the implications of their actions on others, especially with regard to information management (IM). Creating business rules and running audit reports can only do so much. Based upon my nearly fifteen years of working in different IM capacities across the globe, I'd categorize most organizations' related efforts as shown in Figure 1.4.

For every organization currently managing its data very well, many more are doing a poor job. Call it *data dysfunction*, and I'm far from the only one who has noticed this disturbing fact. As for why this is the case, the reasons vary, but I asked my friend Tony Fisher for his take on the matter. Fisher is the founder of DataFlux and the author

Figure 1.4 The Organizational Data Management Pyramid

of *The Data Asset: How Smart Companies Govern Their Data for Business Success*. Fisher told me

> The problem with data management in most organizations today is that they manage their data to support the needs of a specific application. While this may be beneficial in the context of any one application, it falls woefully short in supporting the needs of the entire enterprise. With more sources of data—and more variety of data and larger volumes of data, organizations will continue to struggle if they don't adopt a more contemporary and holistic mind-set. They need to reorient themselves, aligning their data management practices with an *organizational* strategy instead of an application strategy.[11]

In other words, each department in an organization tends to emphasize its own data management and application needs. While this may seem to make sense for each department, on a broader level, this approach ultimately results in a great deal of organizational dysfunction. Many employees, departments, teams, groups, and organizations operate at a suboptimal level. Sadly, they often make routine and even strategic decisions without complete or accurate information. For instance, how can a VP of Sales make accurate sales forecasts when his organization lacks accurate master customer data? How can an HR Director make optimal recruiting decisions without knowing where her company's best and brightest come from? They can't—at least easily.

Far too many organizations struggle just trying to manage their structured data. (These are the ones too busy to even dabble with the other kinds of data discussed earlier in his chapter.) The results can be seen at individual employee levels. Because of poor data management, many employees continue to spend far too much time attempting to answer what should be relatively simple and straightforward business questions. Examples include

- **HR:** How many employees work here? Which skills do our employees have? Which ones are they lacking?
- **Payroll:** Are employees being paid accurately?
- **Finance and Accounting:** Which departments are exceeding their budgets?

- **Sales:** How many products do we sell? Which ones are selling better than others? Which customers are buying from us? How many customers in New York have bought from us in the past year?
- **Supply Chain:** What are our current inventory levels of key products and parts? When can we expect them to be replenished? Will we have enough inventory to meet current and future demand?
- **Marketing:** What's our company's market share? How has that changed from last year or last quarter?

For organizations with cultures that have embraced information-based decision-making, answers to questions like these can be derived quickly.

THE ENTERPRISE AND THE BRAVE NEW BIG DATA WORLD

The questions at the end of the previous section represent vast simplifications of many employees' jobs. I am certainly not implying that the entire responsibilities of employees and departments can be reduced a few bullet points. The previous synopses only serve to illustrate the fact that bad data is downright confounding. Poor data management precludes many organizations and their employees from effectively blocking and tackling. Many cannot easily produce accurate financial reports, lists of customers, and the like. As a result, at these data-challenged companies, accurately answering broader and mission-critical questions like, "How do we sell more products?" is extremely difficult, if not impossible.

On occasion, executives will try to address big-picture items like these by contracting external agencies or perhaps bringing in external consultants. As one of these folks, let me be blunt: we aren't magicians. While we hopefully bring unique perspectives, useful methodologies, and valuable skills to the table, we ultimately face the same data limitations as everyday employees.

The pernicious effects of bad data have plagued organizations for decades, not to mention consultants like me. Inconsistent, duplicate, and incomplete data mystifies everyone. At a minimum, data issues

complicate the answering of both simple and more involved business questions. At worst, they make it impossible to address key issues.

Now let's move to the other end of the spectrum. Consider the relatively few organizations that manage their data exceptionally well (Figure 1.4). I have found that, all else being equal, employees in companies at the top of the pyramid are more productive than those at the bottom. The former group can focus more of its energies on answering bigger, broader questions. They don't have to routinely massage or second-guess organizational data. For instance, it's easier for the head of HR to develop an effective succession plan when the organization knows exactly which employees have which skills. With impeccable customer and sales data, it's more feasible for the chief marketing officer (CMO) to optimize her company's marketing spend.

Today, "simple" and traditional business questions still matter, as do their answers. The advent of Twitter and YouTube certainly did not obviate the need for organizations to effectively manage their structured and transactional data. In fact, doing so remains critical to running a successful enterprise. At the same time, though, Web 2.0 is a game-changer. We're not in Kansas anymore. It is no longer sufficient for organizations to focus exclusively on collecting, analyzing, and managing "table-friendly" data. Twitter, YouTube, Facebook, and their ilk only mean that, for any organization, structured data now only tells part of the story. Texts, tweets, social review sites like Yelp and Angie's List, Facebook likes, Google +1s, photos, blog posts, and viral videos collectively represent a new and important breed of cat. This data cannot be easily (if at all) stored, retrieved, and analyzed via standalone employee databases, large database tables, or often even traditional data warehouses.

The Data Disconnect

Today many organizations suffer a disconnect between new forms of data and old tools that handled, well, old types of data. Many employees cannot even begin to answer critical business questions made more complicated by the Big Data explosion. Examples include

- How do our customers feel about our products or customer service?

- What products would our customers consider buying?
- When is the best time of the year to launch a new product?
- What are people publicly saying about our latest commercial or brand?

Why the disconnect? Several reasons come to mind. First, let's discuss the elephant in the room. Many businesspeople don't think of tweets, blog posts, and Pinterest pins as "data" in the conventional sense, much less potentially valuable data. Why should they waste their time with such nonsense, especially when they continue to struggle with "real" data? Fewer than half of organizations currently collect and analyze data from social media, according to a recent IBM survey.[12] To quote from *Cool Hand Luke,* "Some men you just can't reach."

At least there's some good news: not everyone is in denial over Big Data. Some organizations and employees do get it—and this book will introduce you to many of them. Count among them the U.S. federal government. In March 2012, it formally recognized the power of—and need for—Big Data tools and programs.[13]

But there's no magic "Big Data switch." Simply recognizing that Big Data matters does not mean that organizations can *immediately* take advantage of it, at least in any meaningful way. Many early Big Data zealots suffer from a different problem: they lack the requisite tools to effectively handle Big Data. When it comes to unstructured data, standard reports, ad hoc queries, and even many powerful data warehouses and business intelligence (BI) applications just don't cut it. They were simply not built to house, retrieve, and interpret Big Data. While these old stalwarts are far from dated, they cannot accommodate the vast technological changes of the past seven years—and the data generated by these changes. To paraphrase The Who, the old boss isn't the same as the new boss.

Big Tools and Big Opportunities

As the Chinese say, there is opportunity in chaos. While relatively recent, the rise of Big Data has hardly gone unnoticed. New applications and technologies allow organizations to take advantage of Big Data.

Equipped with these tools, organizations are deepening their understanding of essential business questions.

But, as we'll see in this book, Big Data can do much, much more than answer even complex, predefined questions. Predictive analytics and sentiment analysis are not only providing insights into existing problems, but addressing unforeseen ones. In effect, they are suggesting new and important questions, as well as their answers. Through Big Data, organizations are identifying issues, trends, problems, and opportunities that human beings simply cannot.

Unlocking the full power of Big Data is neither a weekend project nor a hackathon. To be successful here, organizations need to do far more than purchase an expensive new application and hire a team of consultants to deploy it. (In fact, Hadoop, one of today's most popular Big Data tools, is available for free download to anyone who wants it.) Rather, to succeed at Big Data, CXOs need to do several things:

- Recognize that the world has changed—and isn't changing back.
- Disavow themselves of antiquated mind-sets.
- Realize that Big Data represents a big opportunity.
- Understand that existing tools like relational databases are insufficient to handle the explosion of unstructured data.
- Embrace new and Big Data–specific tools—and encourage employees to utilize and experiment with them.

The following chapter will make the compelling business case for organizations to embrace Big Data.

SUMMARY

This chapter has described the evolution of enterprise data and the arrival of the Data Deluge. It has distinguished among the different types of data: structured, semi-structured, and unstructured. With respect to managing their structured data, most organizations in 2013 are doing only passable jobs at best. This squeaking by has rarely come quickly and easily Today, there's a great deal more unstructured data than its structured equivalent (although there's still plenty of the latter).

It's high time for organizations to do more with data beyond just keeping the lights on. There's a big opportunity with Big Data, but

what exactly is it? Answering that question is the purpose of the next chapter. It characterizes Big Data.

NOTES

1. "The Data Deluge," February 25, 2012, www.economist.com/node/15579717, retrieved December 11, 2012.
2. Harris, Jim, "Data Management: The Next Generation," October 24, 2012, www.dataroundtable.com/?p=11582, retrieved December 11, 2012.
3. Rogers, Shawn, "Big Data Is Scaling BI and Analytics," September 1, 2011, www.information-management.com/issues/21_5/big-data-is-scaling-bi-and-analytics-10021093-1.html, retrieved December 11, 2012.
4. Grimes, Seth, "Unstructured Data and the 80 Percent Rule," copyright 2011, http://clarabridge.com/default.aspx?tabid=137&ModuleID=635&ArticleID=551, retrieved December 11, 2012.
5. Pascal, Fabian, "'Unstructured Data': Why This Popular Term Is Really a Contradiction," September 19, 2012, www.allanalytics.com/author.asp?section_id=2386&doc_id=250980, retrieved December 11, 2012.
6. "What to Do About 'Unstructured Data,'" May 15, 2011, www.dbms2.com/2011/05/15/what-to-do-about-unstructured-data/, retrieved December 11, 2012.
7. Reisinger, Don, "Facebook Acquires Face.com for Undisclosed Sum," June 18, 2012, http://news.cnet.com/8301-1023_3-57455287-93/facebook-acquires-face.com-for-undisclosed-sum/, retrieved December 11, 2012.
8. Couts, Andrew, "Lytro: The Camera That Could Change Photography Forever," June 22, 2011, www.digitaltrends.com/photography/lytro-the-camera-that-could-change-photography-forever/, retrieved December 11, 2012.
9. Lacy, Sarah, "Lytro Launches to Transform Photography with $50M in Venture Funds (TCTV)," June 21, 2011, http://techcrunch.com/2011/06/21/lytro-launches-to-transform-photography-with-50m-in-venture-funds-tctv/, retrieved December 11, 2012.
10. Harris, Derrick, "Like Your Data Big? How About 5 *Trillion* Records?," January 4, 2012, http://gigaom.com/cloud/like-your-data-big-how-about-5-trillion-records/, retrieved December 11, 2012.
11. Personal conversation with Fisher, October 25, 2012.
12. Cohan, Peter, "Big Blue's Bet on Big Data," November 1, 2012, www.forbes.com/sites/petercohan/2012/11/01/big-blues-bet-on-big-data, retrieved December 11, 2012.
13. Kalil, Tom, "Big Data Is a Big Deal," March 29, 2012, www.whitehouse.gov/blog/2012/03/29/big-data-big-deal, retrieved December 11, 2012.

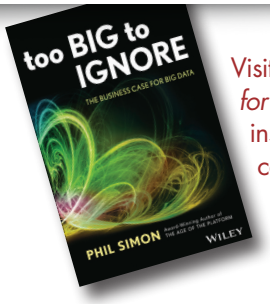
About the Author

Phil Simon is a sought-after keynote speaker and the author of five books, including the award-winning *The Age of the Platform*. While not writing and speaking, he consults organizations on how to optimize their use of technology. His contributions have been featured on NBC, CNBC, The *New York Times*, *Inc.* Magazine, *Bloomberg BusinessWeek*, *The Huffington Post*, *The Globe and Mail*, *Fast Company*, and many other mainstream media outlets. He holds degrees from Carnegie Mellon and Cornell University. You can find him on Twitter at @philsimon, and his home page is www.philsimon.com. He lives in Henderson, Nevada.

ANALYTICS

Know what's hot.

The topic of analytics is on fire right now. With SAS,® you can discover innovative ways to increase profits, reduce risk, predict trends and turn data assets into new business opportunities.



Visit **go.sas.com/toobig** to purchase *Too Big to Ignore: The Business Case for Big Data*. This recommended book offers case studies, examples and insights from industry experts, and discusses tools and applications that can help your business effectively leverage Big Data.

sas®

THE POWER TO KNOW.®