PART
ONE

The Rise of Big Data

CHAPTER 1

What Is Big Data and Why Does It Matter?

Perhaps nothing will have as large an impact on advanced analytics in the coming years as the ongoing explosion of new and powerful data sources. When analyzing customers, for example, the days of relying exclusively on demographics and sales history are past. Virtually every industry has at least one completely new data source coming online soon, if it isn't here already. Some of the data sources apply widely across industries; others are primarily relevant to a very small number of industries or niches. Many of these data sources fall under a new term that is receiving a lot of buzz: big data.

Big data is sprouting up everywhere and using it appropriately will drive competitive advantage. Ignoring big data will put an organization at risk and cause it to fall behind the competition. To stay competitive, it is imperative that organizations aggressively pursue capturing and analyzing these new data sources to gain the insights that they offer. Analytic professionals have a lot of work to do! It won't be easy to incorporate big data alongside all the other data that has been used for analysis for years.

This chapter begins with some background on big data and what it is all about. Then it will cover a number of considerations in terms of how an organization can make use of big data. Readers will need

to understand what is in this chapter as much as or more than anything else in the book if they are to tame the big data tidal wave successfully.

WHAT IS BIG DATA?

There is not a consensus in the marketplace as to how to define big data, but there are a couple of consistent themes. Two sources have done a good job of capturing the essence of what most would agree big data is all about. The first definition is from Gartner's Merv Adrian in a Q1, 2011 *Teradata Magazine* article. He said, "Big data exceeds the reach of commonly used hardware environments and software tools to capture, manage, and process it within a tolerable elapsed time for its user population."[1] Another good definition is from a paper by the McKinsey Global Institute in May 2011: "Big data refers to data sets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze."[2]

These definitions imply that what qualifies as big data will change over time as technology advances. What was big data historically or what is big data today won't be big data tomorrow. This aspect of the definition of big data is one that some people find unsettling. The preceding definitions also imply that what constitutes big data can vary by industry, or even organization, if the tools and technologies in place vary greatly in capability. We will talk more about this later in the chapter in the section titled "Today's Big Data Is Not Tomorrow's Big Data."

A couple of interesting facts in the McKinsey paper help bring into focus how much data is out there today:

- $600 today can buy a disk drive that will store all of the world's music.
- There are 30 billion pieces of information shared on Facebook each month.
- Fifteen of 17 industry sectors in the United States have more data per company on average than the U.S. Library of Congress.[3]

THE "BIG" IN BIG DATA ISN'T JUST ABOUT VOLUME

While big data certainly involves having a lot of data, big data doesn't refer to data volume alone. Big data also has increased velocity (i.e., the rate at which data is transmitted and received), complexity, and variety compared to data sources of the past.

Big data isn't just about the size of the data in terms of how much data there is. According to the Gartner Group, the "big" in big data also refers to several other characteristics of a big data source.[4] These aspects include not just increased volume but increased velocity and increased variety. These factors, of course, lead to extra complexity as well. What this means is that you aren't just getting a lot of data when you work with big data. It's also coming at you fast, it's coming at you in complex formats, and it's coming at you from a variety of sources.

It is easy to see why the wealth of big data coming toward us can be likened to a tidal wave and why taming it will be such a challenge! The analytics techniques, processes, and systems within organizations will be strained up to, or even beyond, their limits. It will be necessary to develop additional analysis techniques and processes utilizing updated technologies and methods in order to analyze and act upon big data effectively. We will talk about all these topics before the book is done with the goal of demonstrating why the effort to tame big data is more than worth it.

IS THE "BIG" PART OR THE "DATA" PART MORE IMPORTANT?

It is already time to take a brief quiz! Stop for a minute and consider the following question before you read on: What is the most important part of the term *big data*? Is it (1) the "big" part, (2) the "data" part, (3) both, or (4) neither? Take a minute to think about it and once you've locked in your answer, proceed to the next paragraph. In the meantime, imagine the "contestants are thinking" music from a game show playing in the background.

Okay, now that you've locked in your answer let's find out if you got the right answer. The answer to the question is choice (4). Neither the "big" part nor the "data" part is the most important part of big data. Not by a long shot. What organizations do with big data is what is most important. The analysis your organization does against big data combined with the actions that are taken to improve your business are what matters.

Having a big source of data does not in and of itself add any value whatsoever. Maybe your data *is* bigger than mine. Who cares? In fact, having any set of data, however big or small it may be, doesn't add any value by itself. Data that is captured but not used for anything is of no more value than some of the old junk stored in an attic or basement. Data is irrelevant without being put into context and put to use. As with any source of data big or small, the power of big data is in what is done with that data. How is it analyzed? What actions are taken based on the findings? How is the data used to make changes to a business?

Reading a lot of the hype around big data, many people are led to believe that just because big data has high volume, velocity, and variety, it is somehow better or more important than other data. This is not true. As we will discuss later in the chapter in the section titled *Most Big Data Doesn't Matter*, many big data sources have a far higher percentage of useless or low-value content than virtually any historical data source. By the time you trim down a big data source to what you actually need, it may not even be so big any more. But that doesn't really matter, because whether it stays big or whether it ends up being small when you're done processing it, the size isn't important. It's what you do with it.

IT ISN'T HOW BIG IT IS. IT'S HOW YOU USE IT!

We're talking about big data of course! Neither the fact that big data is big nor the fact that it is data adds any inherent value. The value is in how you analyze and act upon the data to improve your business.

The first critical point to remember as we start into the book is that big data is both big and it's data. However, that's not what's going

to make it exciting for you and your organization. The exciting part comes from all the new and powerful analytics that will be possible as the data is utilized. We're going to talk about a number of those new analytics as we proceed.

HOW IS BIG DATA DIFFERENT?

There are some important ways that big data is different from traditional data sources. Not every big data source will have every feature that follows, but most big data sources will have several of them.

First, big data is often automatically generated by a machine. Instead of a person being involved in creating new data, it's generated purely by machines in an automated way. If you think about traditional data sources, there was always a person involved. Consider retail or bank transactions, telephone call detail records, product shipments, or invoice payments. All of those involve a person doing something in order for a data record to be generated. Somebody had to deposit money, or make a purchase, or make a phone call, or send a shipment, or make a payment. In each case, there is a person who is taking action as part of the process of new data being created. This is not so for big data in many cases. A lot of sources of big data are generated without any human interaction at all. A sensor embedded in an engine, for example, spits out data about its surroundings even if nobody touches it or asks it to.

Second, big data is typically an entirely new source of data. It is not simply an extended collection of existing data. For example, with the use of the Internet, customers can now execute a transaction with a bank or retailer online. But the transactions they execute are not fundamentally different transactions from what they would have done traditionally. They've simply executed the transactions through a different channel. An organization may capture web transactions, but they are really just more of the same old transactions that have been captured for years. However, actually capturing browsing behaviors as customers execute a transaction creates fundamentally new data which we'll discuss in detail in Chapter 2.

Sometimes "more of the same" can be taken to such an extreme that the data becomes something new. For example, your power meter has probably been read manually each month for years. An

argument can be made that automatic readings every 15 minutes by a Smart Meter is more of the same. It can also be argued that it is so much more of the same and that it enables such a different, more in-depth level of analytics that such data is really a new data source. We'll discuss this data in Chapter 3.

Third, many big data sources are not designed to be friendly. In fact, some of the sources aren't designed at all! Take text streams from a social media site. There is no way to ask users to follow certain standards of grammar, or sentence ordering, or vocabulary. You are going to get what you get when people make a posting. It can be difficult to work with such data at best and very, very ugly at worst. We'll discuss text data in Chapters 3 and 6. Most traditional data sources were designed up-front to be friendly. Systems used to capture transactions, for example, provide data in a clean, preformatted template that makes the data easy to load and use. This was driven in part by the historical need to be highly efficient with space. There was no room for excess fluff.

BIG DATA CAN BE MESSY AND UGLY

Traditional data sources were very tightly defined up-front. Every bit of data had a high level of value or it would not be included. With the cost of storage space becoming almost negligible, big data sources are not always tightly defined up-front and typically capture everything that may be of use. This can lead to having to wade through messy, junk-filled data when doing an analysis.

Last, large swaths of big data streams may not have much value. In fact, much of the data may even be close to worthless. Within a web log, there is information that is very powerful. There is also a lot of information that doesn't have much value at all. It is necessary to weed through and pull out the valuable and relevant pieces. Traditional data sources were defined up-front to be 100 percent relevant. This is because of the scalability limitations that were present. It was far too expensive to have anything included in a data feed that wasn't critical. Not only were data records predefined, but every piece of data in them was high-value. Storage space is no longer a primary con-

straint. This has led to the default with big data being to capture everything possible and worry later about what matters. This ensures nothing will be missed, but also can make the process of analyzing big data more painful.

HOW IS BIG DATA MORE OF THE SAME?

As with any new topic getting a lot of attention, there are all sorts of claims about how big data is going to fundamentally change everything about how analysis is done and how it is used. If you take the time to think about it, however, it really isn't the case. It is an example where the hype is going beyond the reality.

The fact that big data is big and poses scalability issues isn't new. Most new data sources were considered big and difficult when they first came into use. Big data is just the next wave of new, bigger data that pushes current limits. Analysts were able to tame past data sources, given the constraints at the time, and big data will be tamed as well. After all, analysts have been at the forefront of exploring new data sources for a long time. That's going to continue.

Who first started to analyze call detail records within telecom companies? Analysts did. I was doing churn analysis against mainframe tapes at my first job. At the time, the data was mind-boggling big. Who first started digging into retail point-of-sale data to figure out what nuggets it held? Analysts did. Originally, the thought of analyzing data about tens to hundreds of thousands of products across thousands of stores was considered a huge problem. Today, not so much.

The analytical professionals who first dipped their toe into such sources were dealing with what at the time were unthinkably large amounts of data. They had to figure out how to analyze it and make use of it within the constraints in place at the time. Many people doubted it was possible, and some even questioned the value of such data. That sounds a lot like big data today, doesn't it?

Big data really isn't going to change what analytic professionals are trying to do or why they are doing it. Even as some begin to define themselves as data scientists, rather than analysts, the goals and objectives are the same. Certainly the problems addressed will evolve with

big data, just as they have always evolved. But at the end of the day, analysts and data scientists will simply be exploring new and unthinkably large data sets to uncover valuable trends and patterns as they have always done. For the purposes of this book, we'll include both traditional analysts and data scientists under the umbrella term "analytic professionals." We'll also cover these professionals in much more detail in Chapters 7, 8, and 9. The key takeaway here is that the challenge of big data isn't as new as it first sounds.

YOU HAVE NOTHING TO FEAR

In many ways, big data doesn't pose any problems that your organization hasn't faced before. Taming new, large data sources that push the current limits of scalability is an ongoing theme in the world of analytics. Big data is simply the next generation of such data. Analytical professionals are well-versed in dealing with these situations. If your organization has tamed other data sources, it can tame big data, too.

Big data will change some of the tactics analytic professionals use as they do their work. New tools, methods, and technologies will be added alongside traditional analytic tools to help deal more effectively with the flood of big data. Complex filtering algorithms will be developed to siphon off the meaningful pieces from a raw stream of big data. Modeling and forecasting processes will be updated to include big data inputs on top of currently exiting inputs. We'll discuss these topics more in Chapters 4, 5, and 6.

The preceding tactical changes don't fundamentally alter the goals or purpose of analysis, or the analysis process itself. Big data will certainly drive new and innovative analytics, and it will force analytic professionals to continue to get creative within their scalability constraints. Big data will also only get bigger over time. However, incorporating big data really isn't that much different from what analysts have always done. They are ready to meet the challenge.

RISKS OF BIG DATA

Big data does come with risks. One risk is that an organization will be so overwhelmed with big data that it won't make any progress. The

key here, as we will discuss in Chapter 8, is to get the right people involved so that doesn't happen. You need the right people attacking big data and attempting to solve the right kinds of problems. With the right people addressing the right problems, organizations can avoid spinning their wheels and failing to make progress.

Another risk is that costs escalate too fast as too much big data is captured before an organization knows what to do with it. As with anything, avoiding this is a matter of making sure that progress moves at a pace that allows the organization to keep up. It isn't necessary to go for it all at once and capture 100 percent of every new data source starting tomorrow. What is necessary is to start capturing samples of the new data sources to learn about them. Using those initial samples, experimental analysis can be performed to determine what is truly important within each source and how each can be used. Building from that base, an organization will be ready to effectively tackle a data source on a large scale.

Perhaps the biggest risk with many sources of big data is privacy. If everyone in the world was good and honest, then we wouldn't have to worry much about privacy. But everyone is not good and honest. In fact, in addition to individuals, there are also companies that are not good and honest. There are even entire governments that are not good and honest. Big data has the potential to be problematic here. Privacy will need to be addressed with respect to big data, or it may never meet its potential. Without proper restraints, big data has the potential to unleash such a groundswell of protest that some sources of it may be shut down completely.

Consider the attention received by recent security breaches that led to credit card numbers and classified government documents being stolen and posted online. It isn't a stretch to say that if data is being stored, somebody will try and steal it. Once the bad guys get their hands on data they will do bad things with it. There have also been high-profile cases of major organizations getting into trouble for having ambiguous or poorly defined privacy policies. This has led to data being used in ways that consumers didn't understand or support, causing a backlash. Both self-regulation and legal regulation of the uses of big data will need to evolve as the use of big data explodes.

Self-regulation is critical. After all, it shows that an industry cares. Industries should regulate themselves and develop rules that everyone can live with. Self-imposed rules are usually better and less restrictive than those created when a government entity steps in because an industry didn't do a good job of policing itself.

PRIVACY WILL BE A HUGE ISSUE WITH BIG DATA

Given the sensitive nature of many sources of big data, privacy concerns will be a major focal point. Once data exists, dishonest people will try to use it in ways you wouldn't approve of without your consent. Policies and protocols for the handling, storage, and application of big data are going to need to catch up with the analysis capabilities that already exist. Be sure to think through your organization's approach to privacy up front and make your position totally clear and transparent.

People are already concerned about how their web browsing history is tracked. There are also concerns about the tracking of people's locations and actions through cell phone applications and GPS systems. Nefarious uses of big data are possible, and if it is possible someone will try it. Therefore, steps need to be taken to stop that from happening. Organizations will need to clearly explain how they will keep data secure and how they will use it if the general population is going to accept having their data captured and analyzed.

WHY YOU NEED TO TAME BIG DATA

Many organizations have done little, if anything, with big data yet. Luckily, your organization is not too far behind in 2012 if you have ignored big data so far, unless you are in an industry, such as ecommerce, where analyzing big data is already standard. That will change soon, however, as momentum is picking up rapidly. So far, most organizations have missed only the chance to be on the leading edge. That is actually just fine with many organizations. Today, they have a chance to get ahead of the pack. Within a few years, any organization that isn't analyzing big data will be late to the game and will be stuck

playing catch up for years to come. The time to start taming big data is now.

It isn't often that a company can leverage totally new data sources and drive value for its business while the competition isn't doing the same thing. That is the huge opportunity in big data today. You have a chance to get ahead of much of your competition and beat them to the punch. We will continue to see examples in the coming years of businesses transforming themselves with the analysis of big data. Case studies will tell the story about how the competition was left in the dust and caught totally off guard. It is already possible to find compelling results being discussed in articles, at conferences, and elsewhere today. Some of these case studies are from companies in industries considered dull, old, and stodgy. It isn't just the sexy, new industries like ecommerce that are involved. We'll look at a variety of examples of how big data can be used in Chapters 2 and 3.

THE TIME IS NOW!

Your organization needs to start taming big data now. As of today, you've only missed the chance to be on the bleeding edge if you've ignored big data. Today, you can still get ahead of the pack. In a few years, you'll be left behind if you are still sitting on the sidelines. If your organization is already committed to capturing data and using analysis to make decisions, then going after big data isn't a stretch. It is simply an extension of what you are already doing today.

The fact is that the decision to start taming big data shouldn't be a big stretch. Most organizations have already committed to collecting and analyzing data as a core part of their strategy. Data warehousing, reporting, and analysis are ubiquitous. Once an organization has bought into the idea that data has value, then taming and analyzing big data is just an extension of that commitment. Don't let a naysayer tell you it isn't worth exploring big data, or that it isn't proven, or that it's too risky. Those same excuses would have prevented any of the progress made in the past few decades with data and analysis. Focus those who are uncertain or nervous about big data on the fact that big data is simply an extension of what the organization is already

doing. There is nothing earth-shatteringly new and different about it and nothing to fear.

THE STRUCTURE OF BIG DATA

As you read about big data, you will come across a lot of discussion on the concept of data being structured, unstructured, semi-structured, or even multi-structured. Big data is often described as unstructured and traditional data as structured. The lines aren't as clean as such labels suggest, however. Let's explore these three types of data structure from a layman's perspective. Highly technical details are out of scope for this book.

Most traditional data sources are fully in the structured realm. This means traditional data sources come in a clear, predefined format that is specified in detail. There is no variation from the defined formats on a day-to-day or update-to-update basis. For a stock trade, the first field received might be a date in a MM/DD/YYYY format. Next might be an account number in a 12-digit numeric format. Next might be a stock symbol that is a three- to five-digit character field. And so on. Every piece of information included is known ahead of time, comes in a specified format, and occurs in a specified order. This makes it easy to work with.

Unstructured data sources are those that you have little or no control over. You are going to get what you get. Text data, video data, and audio data all fall into this classification. A picture has a format of individual pixels set up in rows, but how those pixels fit together to create the picture seen by an observer is going to vary substantially in each case. There are sources of big data that are truly unstructured such as those preceding. However, most data is at least semi-structured.

Semi-structured data has a logical flow and format to it that can be understood, but the format is not user-friendly. Sometimes semi-structured data is referred to as multi-structured data. There can be a lot of noise or unnecessary data intermixed with the nuggets of high value in such a feed. Reading semi-structured data to analyze it isn't as simple as specifying a fixed file format. To read semi-structured data, it is necessary to employ complex rules that dynamically determine how to proceed after reading each piece of information.

Web logs are a perfect example of semi-structured data. Web logs are pretty ugly when you look at them; however, each piece of information does, in fact, serve a purpose of some sort. Whether any given piece of a web log serves your purposes is another question. See Figure 1.1 for an example of a raw web log.

WHAT STRUCTURE DOES YOUR BIG DATA HAVE?

Many sources of big data are actually semi-structured or multi-structured, not unstructured. Such data does have a logical flow to it that can be understood so that information can be extracted from it for analysis. It just isn't as easy to deal with as traditional structured data sources. Taming semi-structured data is largely a matter of putting in the extra time and effort to figure out the best way to process it.

There is logic to the information in the web log even if it isn't entirely clear at first glance. There are fields, there are delimiters, and there are values just like in a structured source. However, they do not follow each other consistently or in a set way. The log text generated by a click on a web site right now can be longer or shorter than the log text generated by a click from a different page one minute from now. In the end, however, it's important to understand that semi-structured data does have an underlying logic. It is possible to develop relationships between various pieces of it. It simply takes more effort than structured data.

Raw Web Log Data

96.255.99.50 - - [01/Jun/2010:05:28:07 +0000] "GET /origin-log.enquisite.com/d.js?id=a1a3af-ly6l645&referrer=http://www.google.com/search?hl=en&q=budget+planner&aq=5&aqi=g10&aql=&oq=budget+&gs_rfai=&location=https://money.strands.com/content/simple-and-free-monthly-budget-planner&ua=Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 6.0; SLCC1; .NET CLR 2.0.50727; .NET CLR 3.0.30618; .NET CLR 3.5.30729; InfoPath.2)&pc=pgys63w0xgn102in8ms37wka8quxe74e&sc=cr1kto0wmxqik1wlr9p9weh6yxy8q8sa&r=0.07550191624904945 HTTP/1.1" 200 380 "-" "Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 6.0; SLCC1; .NET CLR 2.0.50727; .NET CLR 3.0.30618; .NET CLR 3.5.30729; InfoPath.2)" "ac=bd76aad174480000679a044cfda00e005b130000"

Figure 1.1 Example of a Raw Web Log

Analytic professionals will be more intimidated by truly unstructured data than by semi-structured data. They may have to wrestle with semi-structured data to bend it to their will, but they can do it. Analysts can get semi-structured data into a form that is well structured and can incorporate it into their analytical processes. Truly unstructured data can be much harder to tame and will remain a challenge for organizations even as they tame semi-structured data.

EXPLORING BIG DATA

Getting started with big data isn't difficult. Simply collect some big data and let your organization's analytics team start exploring what it offers. It isn't necessary for an organization to design a production-quality, ongoing data feed to start. It just needs to get the analytics team's hands and tools on some of the data so that exploratory analysis can begin. This is what analysts and data scientists do.

There is an old rule of thumb that 70 to 80 percent of the time developing an analysis is spent gathering and preparing data and only 20 to 30 percent is spent analyzing it. Expect that those guidelines will be low when initially working with big data. Analytic professionals will initially probably spend at least 95 percent, if not close to 100 percent, of their time just figuring out a big data source before they can even think about doing in-depth analysis with it.

It is important to understand that that's okay. Figuring out what a data source is all about is an important part of the analysis process. It may not be glamorous or exciting, but iteratively loading data, examining what it looks like, and adjusting the load processes in order to better target the data that is needed are immensely important. Without completing those steps, it won't be possible to proceed to the analysis phase.

The process of identifying the pieces of big data that contain value and determining how best to extract those pieces accurately is critical. Expect this to take time and do not get frustrated if it takes longer than anticipated. As the process of figuring out the new data sources progresses, analytic professionals and their business sponsors should look for ways to deliver small, quick wins. Showing the organization something of value, no matter how small, will keep people interested

in the process and help them see that there is progress being made. A cross-functional team can't get started and a year later claim they are still figuring out how to do something with big data. It is necessary to come up with some ideas, even if they are small, and make something happen quickly.

DELIVER VALUE AS YOU GO

It will take a lot of effort to figure out how to apply a source of big data to your business. An organization's analytic professionals and their business sponsors must be sure to look for ways to deliver small, quick wins as they go. It will demonstrate to the organization that progress is being made and will build support for further efforts. Such wins can also generate a solid return on investment.

A great example comes from a European retailer. The company wanted to start leveraging detailed web log data. As they built complex, long-term processes to capture the data, they first put in place a few simple processes. They started by identifying what products each customer browsed. The browsing information was used for a basic follow-up e-mail campaign that sent each customer a message if they left the site after viewing products, but did not end up purchasing them. This simple exercise generated a huge amount of revenue for the organization.

Along with a handful of other similarly basic first steps, the company paid for the entire long-term effort of capturing and loading the web data. More important, they hadn't even gotten fancy or dealt with the entire data stream yet. Imagine the returns they are going to see as they proceed to more deeply analyze the data in the future. Due to the quick and early wins, everyone in the organization is excited to continue because they have seen how powerful even the first, basic uses of the data were. Plus, the further efforts are already paid for!

MOST BIG DATA DOESN'T MATTER

The fact is that most big data just doesn't matter. That sure sounds harsh, doesn't it? But it's not meant to be. As we have already

discussed, a big data stream is going to be large in terms of volume, velocity, variety, and complexity. Much of the content of a big data stream won't matter for any given purpose, and some of it won't matter much at all. Taming the big data tidal wave isn't about getting all the water from the wave nicely controlled in a swimming pool. It's more like sipping water from a hose: You slurp out just what you need and let the rest run by.

Within a big data feed there will be some information that has long-term strategic value, some that will be useful only for immediate and tactical use, and some data that won't be used for anything. A key part of taming big data is to determine which pieces fall into which category.

A great example of this is related to the radio frequency identification (RFID) tags we discuss in Chapter 3 that are being placed today on pallets of products when they are shipped. For expensive items, tags are even being placed on individual items. Eventually, tagging is going to move to individual items as the rule rather than the exception. Today, that is cost prohibitive in most cases, so the tags are often placed on each pallet. The tags make it easier to track where the pallets are, when they are loaded and unloaded, and where they are stored.

Imagine a warehouse with tens of thousands of pallets. Each pallet has an RFID tag. RFID readers query the warehouse every 10 seconds saying, basically, "Who is out there?" Each of the pallets responds back: "I am." Let's discuss how in this case big data starts to be narrowed down very quickly.

A pallet first arrives today and first chimes in: "This is pallet 123456789. I'm here." Every 10 seconds over the next three weeks while that pallet is in the warehouse, it is going to reply again and again: "I'm here. I'm here. I'm here." Upon completion of each of the 10 second polls taken, it's very well worth the effort to parse through all the replies and identify any pallets that have had a change in status. This way, it can be validated that any changes were expected, and action can be taken if a pallet changed status unexpectedly.

Once a pallet actually leaves the warehouse, it is no longer responding. After validating that the pallet was expected to leave when it did, all of the intermediate "I'm here" records really don't matter. Over time, all that really matters are the date and time when

the pallet entered the warehouse and the date and time when it left. If those times are three weeks apart, it only makes sense to keep the two time stamps associated with the entry and exit of the pallet. All of the responses at 10 second intervals in between, saying "I'm here. I'm here. I'm here," have no long-term value whatsoever, but it was necessary to collect them. It was necessary to analyze each at the moment it was generated. But the responses outside the first and last have no long-term value. They can be safely thrown away once the pallet is gone.

GET READY TO THROW DATA AWAY

One key to taming big data will be to identify what pieces matter. There will be pieces that have long-term strategic use, pieces that have short-term tactical use, and pieces that don't matter at all. It will seem odd to let a lot of data slip past, but that is par for the course with big data. Throwing data away will take some getting used to.

If raw big data feeds can be kept available for a period of time, this will provide the capability to go back and extract additional data missed when it was first processed. One good example of this is the way that web activity tracking is done today. Most web sites use what is known as a tag-based methodology. With a tag-based methodology, it is necessary to identify up front what text, images, or links it is desired to track users' interactions with. The tags, which are not seen by the user, will report back that a user has done something. Since only tagged items are reported, most browsing information is ignored from the start. The problem is if a request to have a new promotional image tagged is inadvertently missed, there will be no ability to go back and analyze interactions with that image. It has to be tagged before a user browses. It is possible to add a tag later, but only activity from that point forward will be captured.

There are newer methodologies that will parse through raw web logs and enable the identification of anything that occurred without having predefined it. These methods are log-based since they leverage a raw web log directly. The value of this is that if you realize later that you forgot to capture interactions with the promotion image, you can

go and parse through the data again and pull it out. In this case, nothing is thrown away up front, but what to keep will need to be determined at the time of analysis. That is an important capability and is why keeping some historical big data, as long as it is cost-effective to do so, makes sense. How much historical data can be kept will depend on the size of a data feed and how much storage is reasonably available. It is a good idea to leave as much flexibility as possible within those constraints by keeping as much history as available storage will economically allow.

FILTERING BIG DATA EFFECTIVELY

The biggest challenge with big data may not be the analytics you do with it, but the extract, transform, and load (ETL) processes you have to build to get it ready for analysis. ETL is the process of taking a raw feed of data, reading it, and producing a usable set of output. The data is extracted (E) from whatever source it is starting from. The data is next transformed (T) through various aggregations, functions, and combinations to get it into a usable state. Last, the data is loaded (L) into whatever environment will be leveraged to analyze the data. That is the process of ETL.

Let's go back to the analogy we discussed earlier: sipping water out of a hose. When you're drinking water out of a hose, you don't really care which parts of the stream of water get in your mouth. With big data, you care very much about which parts of the data stream get captured. It will be necessary to explore and understand the entire data stream first. Only then can you filter down to the pieces that you need. This is why the up-front effort to tame big data can take so long.

SIPPING FROM THE HOSE

Working with big data is a lot like taking a drink from a hose. Most of the data will run past, just like most of the water does. The goal is to sip the right amount of data out of the data stream as it flows past, not to try and drink it all. By focusing on the important pieces of the data, it makes big data easier to handle and keeps efforts focused on what is important.

Analytic processes may require filters on the front end to remove portions of a big data stream when it first arrives. There will be other filters along the way as the data is processed. For example, when working with a web log, a rule might be to filter out up front any information on browser versions or operating systems. Such data is rarely needed except for operational reasons. Later in the process, the data may be filtered to specific pages or user actions that need to be examined for the business issues to be addressed.

The complexity of the rules and the magnitude of the data being removed or kept at each stage will vary by data source and by business problem. The load processes and filters that are put on top of big data are absolutely critical. Without getting those correct, it will be very difficult to succeed. Traditional structured data doesn't require as much effort in these areas since it is specified, understood, and standardized in advance. With big data, it is necessary to specify, understand, and standardize it as part of the analysis process in many cases.

MIXING BIG DATA WITH TRADITIONAL DATA

Perhaps the most exciting thing about big data isn't what it will do for a business by itself. It's what it will do for a business when combined with an organization's other data.

Browsing history, for example, is very powerful. Knowing how valuable a customer is and what they have bought in the past across all channels makes web data even more powerful by putting it in a larger context. We'll explore this in detail in Chapter 2.

Smart-grid data is very powerful for a utility company. Knowing the historical billing patterns of customers, their dwelling type, and other factors makes data from a smart meter even more powerful by putting it in a larger context. We'll look at this in Chapter 3.

The text from customer service online chats and e-mails is powerful. Knowing the detailed product specifications of the products being discussed, the sales data related to those products, and historical product defect information makes that text data even more powerful by putting it in a larger context. We'll explore this topic from different perspectives in Chapters 3 and 6.

A large part of the reason why Enterprise Data Warehouses (EDWs) have become such a widespread corporate tool isn't to centralize a bunch of data marts to save hardware and software costs. An EDW adds value by allowing different data sources to intermix and enhance one another. With an EDW, it is possible to analyze customer and employee data together since they are in one location. They are no longer completely separate. For example, do certain employees increase customer value through their interactions more than others? Such questions are much easier to answer if the data is all in one place. As big data is added in, it just continues to evolve the number and magnitude of problems that can be addressed as ever more types of data can be combined together to add new perspectives and contexts.

MIX IT UP!

The biggest value in big data can be driven by combining big data with other corporate data. By putting what is found in big data in a larger context, the quantity and quality of insights will increase exponentially. This is why big data needs to be folded into an overall data strategy as opposed to having a stand-alone big data strategy.

This is why it is critically important that organizations don't develop a big data strategy that is distinct from their traditional data strategy. That will fail. Big data and traditional data are both pieces of the overall strategy. To succeed, organizations need to develop a cohesive strategy where big data isn't a distinct, standalone concept. Rather, big data must be simply another facet of an enterprise data strategy. From the start, it is necessary to think through and plan not just how to capture and analyze big data by itself, but also how to use it in combination with other corporate data and as a component of a more holistic approach to corporate data.

THE NEED FOR STANDARDS

Will big data continue to be a wild west of crazy formats, unconstrained streams, and lack of definition? Probably not. Over time,

standards will be developed. Many semi-structured data sources will become more structured over time, and individual organizations will fine-tune their big data feeds to be friendlier for analysis. But more important, there will be a move toward industry standards. While text data like e-mail or social media commentary can't be controlled very much on the input end, it *is* possible to standardize the approaches to interpreting such data and using it for analytics. This is already starting to happen.

For example, what words are "good" and what words are "bad"? What contexts exist where the default rules don't apply? Which e-mails are worth exhaustive parsing and analysis, and which can be processed minimally? Standards for the production and generation of big data will develop, as will the standards for the processing and analysis of big data. Both the input and output sides will be addressed. As a result, life will get easier for those tasked with taming it. It will take time and many of the standards that develop will be more of a set of commonly accepted best practices among practitioners than formally stated rules or policies from an official standards organization. Nevertheless, standardization will increase.

STANDARDIZE TO THE EXTENT POSSIBLE

While text data like e-mail can't be controlled very much on the input end, it is possible to standardize the approaches to interpreting such data and using it for analytics. You won't be able to standardize everything about big data, but you can standardize enough to make life much easier. Focus on standardizing the use of big data as much as on standardizing the input feed itself.

Organizations that are quick to embrace big data will have the ability to define and influence the developing standards and therefore be sure that their specific needs are met. Some industries are even getting ahead of the curve. There has been a lot of work to define the parameters of smart-grid data within the utility industry even before the ability to collect the data is in place. By starting out with formal definitions and guidelines, smart-grid data will be much more manageable than if every utility had just started creating data in their own way without thinking it through ahead of time with their peers.

TODAY'S BIG DATA IS NOT TOMORROW'S BIG DATA

As discussed at the start of the chapter, the accepted definitions of big data are somewhat "squishy." There is no specific, universal definition in terms of what qualifies as big data. Rather, big data is defined in relative terms tied to available technology and resources. As a result, what counts as big data to one company or industry may not count as big data to another. A large e-commerce company is going to have a much "bigger" definition of big data than a small manufacturer will.

More important, what qualifies as big data will necessarily change over time as the tools and techniques to handle it evolve alongside raw storage size and processing power. Household demographic files with hundreds of fields and millions of customers were huge and tough to manage a decade or two ago. Now such data fits on a thumb drive and can be analyzed by a low-end laptop. As what qualifies as high volume, high velocity, high variety, and high complexity evolves, so will big data.

"BIG" WILL CHANGE

What's big data today won't be considered big data tomorrow any more than what was considered big a decade ago is considered big today. Big data will continue to evolve. What is impossible or unthinkable today in terms of data volume, velocity, variety, and complexity won't be so years down the road. That's how it has always been, and it will continue as such in the era of big data.

Transactional data in the retail, telecommunications, and banking industries were very big and hard to handle even a decade ago. In fact, such data wasn't widely available for analytics and reporting in many organizations in the late 1990s. Today, such data is considered a necessary and fundamental asset. Virtually every company of any size has access to it.

Similarly, what we are intimidated by today won't be so scary a few years down the road. Clickstream data from the web may be a standard, easily handled data source in 10 years. Actively processing every e-mail, customer service chat, and social media comment may become a standard practice for most organizations. The tracking of

hundreds of metrics per second from an engine may not make anyone break a sweat.

As we tame the current generation of big data streams, other even bigger data sources are going to come along and take their place. What will they be? Nobody has all the answers today. However, following are a few ideas on how current data sources can be upgraded to another magnitude of "big" pretty quickly:

- Imagine web browsing data that expands to include millisecond-level eyeball and mouse movement so that every tiny detail of a user's navigation is captured, instead of just what was clicked on. This is another order of big.
- Imagine video game telemetry data being upgraded to go beyond every button pressed or movement made. Imagine it also containing the eye and body movement of the player along with the location and status of every single object within the scene being played instead of just the objects that are interacted with. That gets massive fast.
- Imagine RFID information being available for every single individual item in every single store, distribution facility, and manufacturing plant globally. Imagine those chips evolving to capture dozens of metrics per second, such as temperatures, humidity, speed, acceleration, pressure, and more. The volume of such data is unthinkable today.
- Imagine capturing and translating to text every conversation anyone has with a customer service or sales line. Add to that all the associated e-mails, online chats, and comments from places such as social media sites or product review sites. Now, go parse, combine, and analyze all of that text. Is your head exploding yet?

The point is that big data is here to stay. Though what we find intimidating today won't be what we find intimidating a few years from now, some new data source will be intimidating. Organizations will need to continue to adjust their methods and their goals to make use of the data as it evolves. Your organization can't adjust and update its methods for handling big data until it has some methods in place, however. So you need to get started!

WRAP-UP

The most important lessons to take away from this chapter are:

- Big data is often defined as data that exceeds the capability of commonly used hardware environments and software tools to capture, manage, and process it within a tolerable elapsed time for its user population.
- Big data is big not just in terms of volume, but also in terms of variety, velocity, and complexity.
- The power of big data is in the analysis you do and the actions you take. It is in neither the "big" part nor the "data" part.
- Big data is often automatically generated by a machine of some sort and is usually not in a user-friendly format. The default is to capture everything and worry about what matters later.
- Big data is just the next wave of new, bigger data that pushes today's limits. From an analytics perspective it isn't any different from past data sources that were large and difficult to handle when they first became available.
- Big data will change some of the tactics and tools that analytic professionals utilize, but it won't fundamentally change why analytics are done or how the value of analytics is assessed.
- Many big data sources are semi-structured. There is logic to a semi-structured data feed, but it may not be pretty. Big data can also be unstructured. In some cases, it is even structured like traditional data sources.
- The biggest risks of big data are the privacy implications that some of the data sources involve. Both self-regulation and legal regulation will be needed as the use of big data evolves.
- Taming the big data tidal wave isn't about controlling all of the data. It is like sipping from a hose. Just skim off the important pieces.
- The most exciting thing about big data is what it will do for a business when combined with other data.
- Big data and traditional data are both pieces of an overall data and analytics strategy. Don't develop a big data strategy that is distinct from your traditional data strategy.

- Big data will continue to evolve. What we think is big and intimidating today won't raise an eyebrow in a decade, but another new data source will!

NOTES

1. Merv Adrian, "Big Data," *Teradata Magazine*, 1:11, www.teradatamagazine.com/v11n01/Features/Big-Data/.
2. McKinsey Global Institute, *Big Data: The Next Frontier for Innovation, Competition, and Productivity*, May 2011.
3. Ibid.
4. *CEO Advisory: "Big Data" Equals Big Opportunity*, Gartner, March 31, 2011.