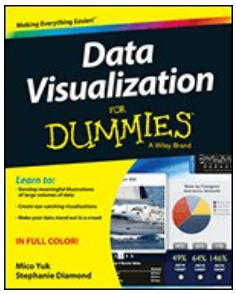


# Data Visualization for Dummies

by Mico Yuk and Stephanie Diamond

John Wiley & Sons (US). (c) 2014. Copying Prohibited.

Reprinted for YI LIN, CVS Caremark

yi.lin@cvscaremark.com

Reprinted with permission as a subscription benefit of **Books24x7**,
http://www.books24x7.com/

All rights reserved. Reproduction and/or distribution in whole or in part in electronic,paper or other forms without written permission is prohibited.

# Chapter 3: Knowing What You Must About Big Data

## In This Chapter

- Understanding Big Data
- Seeing how Big Data is used in business
- Choosing tools for visualization

Anyone who spends time online instinctively knows that data is continuing to grow at a rapid rate. But the magnitude of data growth may shock you. According to Marcia Conner, in her article "Data on Big Data," 90 percent of all the data in the world has been created in the last two years. (See the article at http://marciaconner.com/blog/data-on-big-data.)


If you're doing business today, it's no secret that you're bombarded by all kinds of data. The key is to use this data as an opportunity to create a competitive advantage. In other words, if you take steps to harness and analyze the data you have, you will be ahead of many of your competitors who are still waiting to figure out what to do with their Big Data.

## Defining Big Data

Suddenly, everyone is talking about the presence of Big Data and its impact. Hasn't there always been a steady stream of data that companies can employ? What makes the data so "big" that its mere presence can't be managed?

These questions and many others probably come to mind. The best way to answer them is to think about how you currently interact with devices throughout your day:

- Communicate via smartphones
- Send reports, proposals, and contracts digitally
- Buy products via online accounts
- Use credit cards at retail stores
- Monitor your car and house with various electronic devices such as temperature gauges and dashboard computers

**Remember** Don't forget that data is generated when you interact on social media and in search engines. Most of this data was unavailable a mere 20 years ago.

Every one of the preceding activities throws off mountains of data. The retailer sees what you buy; the smartphone provider knows who you call; the car dashboard records your driving habits. You (and your customers) are a walking mass of data. All this acquired information is what we're talking about when we say *Big Data*.

## Seeing How Big Data Changes Business

It's interesting to note that Big Data changes the way employees work. Here are a few examples:

- Salespeople can access data and make decisions about offers right from the customer's location.
- Customer service representatives can access information about buying habits and needs while they're online or on the phone with customers.


Reprinted for TJGH4\1151844, CVS Caremark John Wiley & Sons (US), John Wiley & Sons, Inc. (c) 2014, Copying Prohibited

- Financial managers can access specific data to meet their needs, and marketing departments can drill down to specific campaign details.

We're sure that you can see the value Big Data brings to any organization that makes an effort to use it. You get not only productivity gains but also faster insight into your problems. Interestingly, most CEOs used to believe that their fortunes were strictly in their customer lists. Now, many are coming to believe that the inclusion of Big Data and the attendant insights it brings will be the source of their fortunes.

With Big Data, you can become customer-centric in ways you never could before. You can go beyond transactional reports about customer buying habits and drill down to customer sentiment and motivation. This chapter looks at what Big Data is and how you can harness it to create data visualizations. The opportunities are limitless.

**Tip** One thing that businesses should focus on is the fact that they can now get answers to questions they couldn't previously ask, which makes asking the right questions a very important factor in Big Data analysis. You're only as good as the questions you ask.

## Getting to know your customers

Companies today have the opportunity to "know" each customer's individual habits. Instead of creating artificial groups of customers who may have some similar interests, businesses can effectively segment and customize their offers to each customer. By knowing customers' habits, companies can not only target niches more accurately but also uncover new markets for their goods, understand customer needs more fully, and extend well-timed special offers.

Here are three major benefits of using Big Data to understand your customers:

- **Customer satisfaction:** You can create fans who are loyal and who speak about your products to others on social media. This situation creates a sales force that money can't buy.
- **More money per customer:** Offering just the right product at the right time increases buying.
- **Retention:** You keep customers longer by showing them that you understand their needs and can provide them with products that meet those needs.

Companies that use Big Data can extract value in several new ways. The following key factors shed light on why companies believe that Big Data presents such a great opportunity:

- **New tools are available**. You can use the tools that are now available to analyze this new data in near real time. These tools include data sensors that collect massive amounts of information about how machines and electronics are performing.
- **New insights are possible**. Being able to analyze the data provides the capability to extract major insights that can increase profits.
- **New types of data are being generated**. Some of social data deals with customer sentiment, which is invaluable to companies. Customer sentiment data lets companies know what offers will succeed because they learn how customers talk about and review their products.
- **New developments make cloud storage a reality**. Warehousing data is less an issue now that businesses can use cloud storage, which is cheaper and just as accessible.

All this sounds pretty good. But Big Data also presents a few challenges. The next section looks at the essence of the problem.

## Discovering the Four *V's*

To understand the challenges of Big Data, the Business Intelligence (BI) community commonly designates what it calls the *Four V's:*

- **Volume:** Tons of data is generated around the globe, and the volume will continue to increase exponentially in the coming years.
- **Variety:** Different and new kinds of structured and unstructured data (such as social media data) are being created. (For more information, see "Collecting structured and unstructured data," later in this chapter.)
- **Velocity:** The pace at which data is mounting is accelerating, and companies' ability to analyze it in real time is crucial to the development of tangible offers that turn shoppers into actual customers.
- **Veracity:** *Veracity* refers to the trustworthiness of data. Is the data that's being generated valid? When analyzing data about customer sentiment on social media, for example, can you trust everything that's been written?

IBM has created an excellent infographic illustrating the Four V's, as shown in Figure 3-1. (You can find the image at http://cdn.dashburst.com/wp-content/uploads/2013/07/ the-four-v-s-of-big-data.jpg.)



Reprinted for TJGH4\1151844, CVS Caremark John Wiley & Sons (US), John Wiley & Sons, Inc. (c) 2014, Copying Prohibited

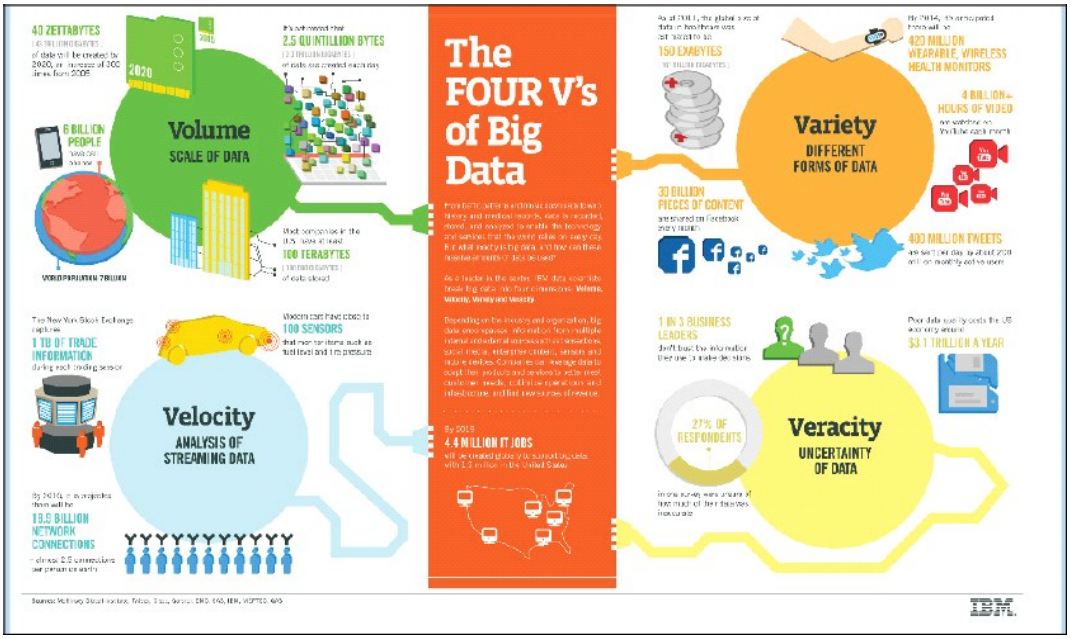


**Figure 3-1:** The Four V ' s.

These four factors influence the management of Big Data. Whether you're a small-business owner or an executive of a major corporation, you are affected by the onslaught of data. The following sections dig a bit deeper into the kinds of data you will be working with.

## Collecting structured and unstructured data

Data collected in the past was structured and could fit into neat rows and columns. An example of this would be an Excel spreadsheet with delimited data (data that was separated by a specific character, such as a comma). Most internal information specialists were content to display this data (such as customer records) in long spreadsheets. They were tasked with reporting what the data said, and everyone used the same results. There was no opportunity to visualize the story that the data told to extract valuable insights. The data wasn't interactive and didn't allow for customization. It was valuable to a point, but there was no way to understand what the customer thought about the product after they bought it. You would only know *that* the product had been bought. And that data is only one part of the puzzle.

Today, companies are facing a mountain of a new type of data: unstructured data, which doesn't always come in a neat package. Following are a few examples of this type of data:

- **Opinions:** Opinions are gathered by review sites such as Yelp, shown in Figure 3-2. You can access the reviews directly or use a tool that scrapes the data from the site so that you can put that data in your own data-viz tool.

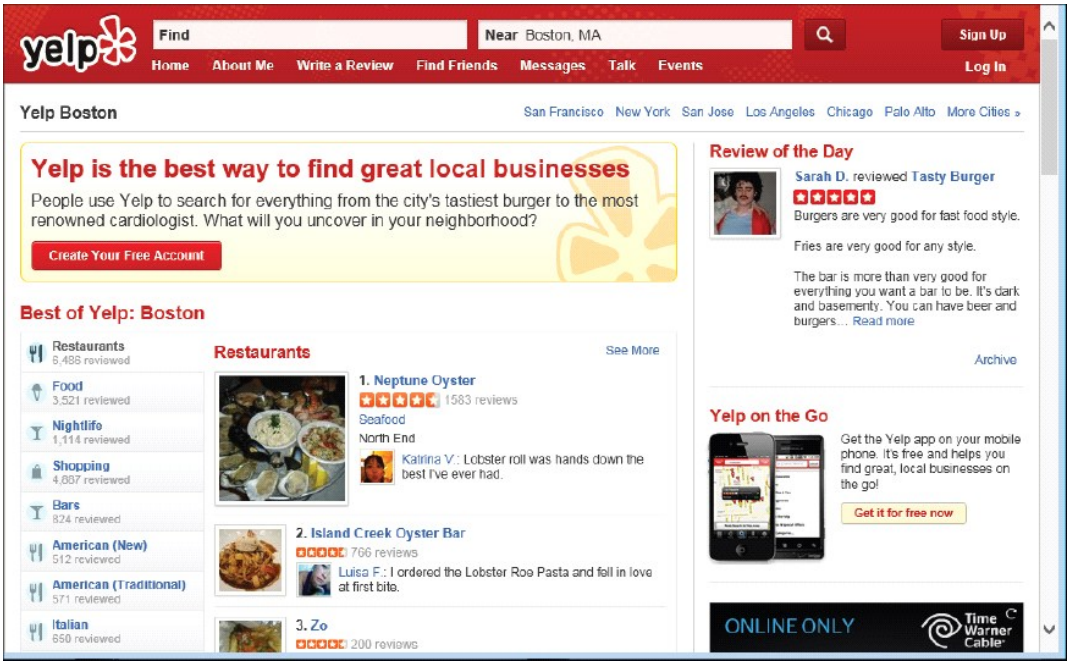


**Figure 3-2:** Restaurant reviews on Yelp.



Reprinted for TJGH4\1151844, CVS Caremark John Wiley & Sons (US), John Wiley & Sons, Inc. (c) 2014, Copying Prohibited

- **Visuals:** Visuals are chosen by users of sites such as Pinterest, shown in Figure 3-3. In the case of Pinterest, you can access the site to see what images about and by your company have been pinned by customers who are searching for your company's name. You may have data about which of your company's pins are being re-pinned by others as well as data about people who have seen your company's product or image elsewhere on the web and have pinned it directly to Pinterest for others to find.

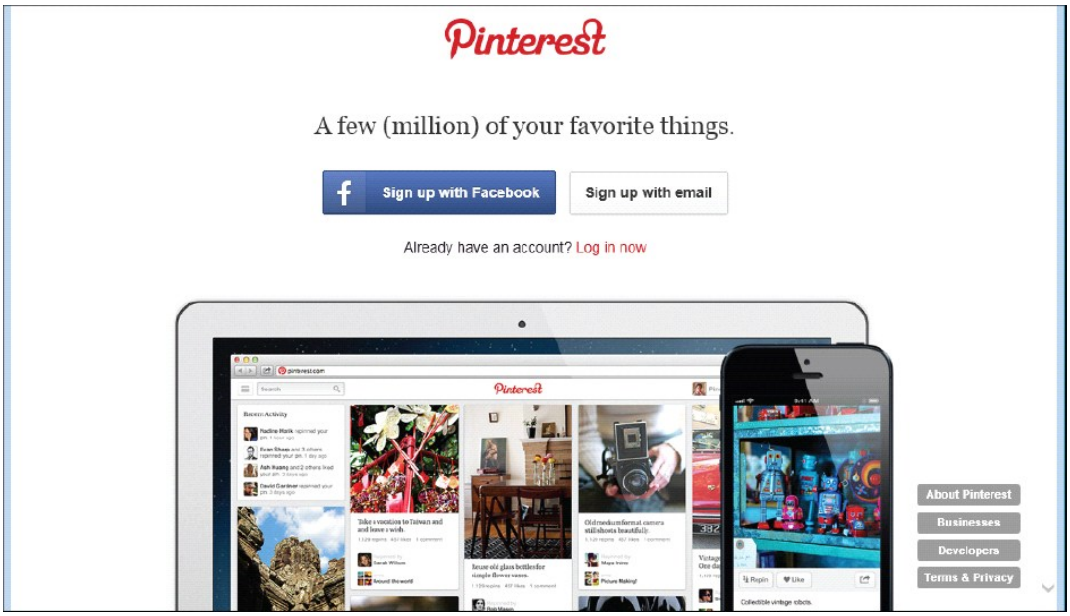


**Figure 3-3:** The Pinterest site can generate a lot of data for a company.

- **Smartphone data:** Phone records, e-mails, and other search data are available from your phone.

This unstructured content represents data that's incredibly valuable to any online business. The key to using the data is to utilize software programs (such SAP) that enable you to combine structured data with unstructured data to gain a greater understanding of the business and its customers. From this analysis, companies can begin to make predictions about customer behavior and revenue generation.

**Technical Stuff** Typically, organizations that use unstructured data use natural language processing software to analyze it.

## Ensuring the use of quality data

The quality of your data, not your choice of tool, determines the value of your visualization. In a 2013 article in the *Harvard Business Review* titled "When Data Visualization Works — and When It Doesn't," author Jim Stikeleather pointed out three elements that affect the efficacy of data:

- **Data quality:** Obviously, if your data is incomplete or full of errors, your data visualization will be useless. But it's not always easy to determine what data is missing and, therefore, how reliable the predictions you make with it will be. It's important to pay attention to the quality of your data up front to make sure that your conclusions are usable. Work with your IT department and major stakeholders to determine as much about your data as you can. Find out about its origins and how often it is updated.
- **Context:** Context refers to your ability to draw conclusions from your data. If you don't understand how the data was sourced, how current it is, and so on, you risk drawing faulty conclusions from it.
- **Biases:** It's important to acknowledge any biases you have about the data before you create your visualization. Do you want the conclusions to support a pet theory? Are you making the data visualization look a certain way to support your conclusions? You must divest yourself of these notions before you begin.

**Warning!** Regarding biases, when you look at any data visualization, it's a good idea to ask yourself whether the data was created by someone who may have a stake in a certain outcome. Sometimes, the developer's bias may be unconscious. Make an agreement with the major stakeholders that the data you use must be certified by IT so that you avoid any bias that might be introduced when the stakeholders themselves provide the data.

## Avoiding Dying by Tool Choice

Throughout this book, we state that the most important thing to focus on is data preparation, not tools. That said, we understand that it's easy to be distracted by the variety of tools that are available for creating visualization.

All tools are not created equal. Tying yourself to a tool before you know what needs to be displayed is a recipe for failure. If your company has already chosen a specific tool, obviously, you have to figure out how to work with that one. But if you have the freedom to decide on a tool, make sure to let the data drive your choice.



Reprinted for TJGH4\1151844, CVS Caremark John Wiley & Sons (US), John Wiley & Sons, Inc. (c) 2014, Copying Prohibited

When choosing a tool, it's important to know what separates truly advanced tools from simpler ones. In a 2012 Forrester Wave report titled "Advanced Data Visualization (ADV) Platforms," Bob Evelson and Noel Yuhanna outlined several points that separate the current crop of advanced tools from older, less powerful tools:

- **Dynamic data content:** The data is interactive and can be updated regularly to show changes.
- **Visual querying:** Users can click icons and other visuals to update the data.
- **Multiple-dimension, linked visualization:** Multiple types of data can be linked to show different dimensions.
- **Animated visualization:** An animated visualization enables you to quickly go to the data you want to see so that you don't have to spend time clicking through data that's not relevant to you.
- **Personalization:** The software assigns different levels of access to the data as well as access to different slices of that data based on the particular user.
- **Business-actionable alerts:** The software triggers alerts that can notify various stakeholders when appropriate.

You may want to evaluate your tools based on these measures. You can access the full report by downloading the PDF at www.sas.com/news/analysts/Forrester_Wave_Advanced_Data_ Visualization_Platforms_Q3_2012.pdf.

We can't know which tool is right for you, of course. Tableau and QlikView have growing adoption rates, so we discuss them in more detail in the following sections.

## Tableau

Tableau has gained popularity because it's an easy-to-use, drag-and-drop software tool. Both free and fee-based versions of the software are available.

The free version (www.tableausoftware.com/public), shown in Figure 3-4, allows users to craft public data to tell a story. Typically, it's used by individuals or small companies. It's optimized for the iPad and Android devices.


**Figure 3-4:** Tableau for Enterprise.

The fee-based version (www.tableausoftware.com) is for larger organizations that want to plug in their own confidential data. A free trial period is available for this version.

### QlikView

QlikView (http://qlikview.com; see Figure 3-5) is relatively new software that has gained favor quickly. It also has free and fee-based versions. It is known for its ease of use and wide range of tools. It can be used by both small and large organizations.



Reprinted for TJGH4\1151844, CVS Caremark John Wiley & Sons (US), John Wiley & Sons, Inc. (c) 2014, Copying Prohibited

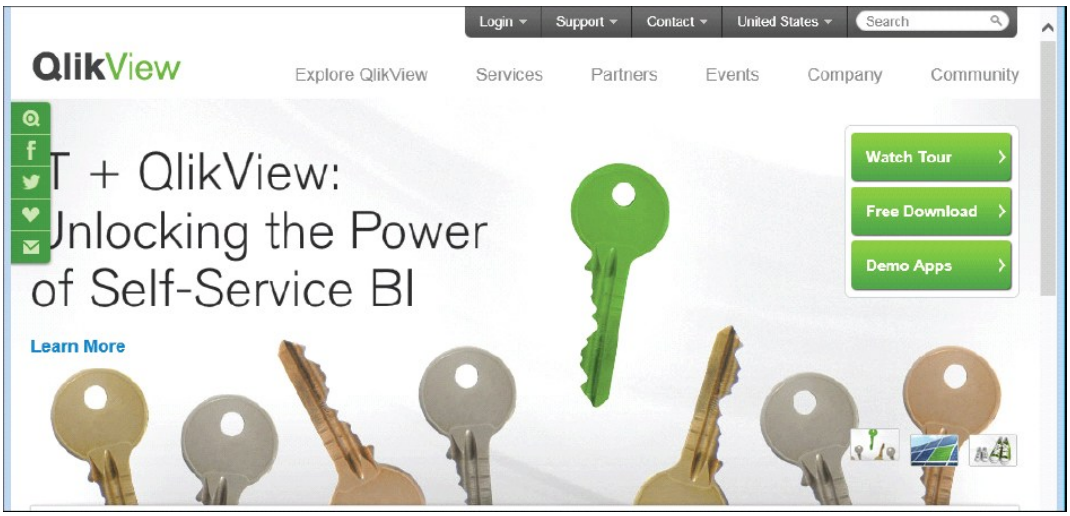


**Figure 3-5:** The QlikView homepage.



Reprinted for TJGH4\1151844, CVS Caremark John Wiley & Sons (US), John Wiley & Sons, Inc. (c) 2014, Copying Prohibited