

**Acquia**<sup>®</sup> THINK AHEAD.

# Are You Ready for Big Data?

An examination of Big Data and its role in  
the next generation digital experience

By DC Denison



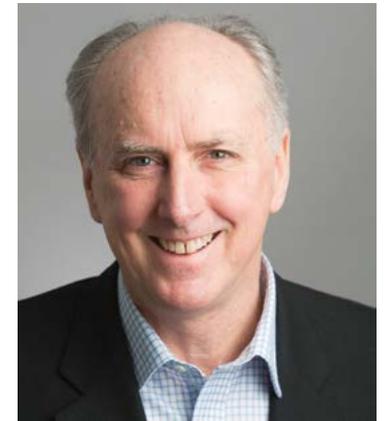
# Table of Contents

Introduction	4
What is Big Data?	4
Should You Start a Big Data Project?	6
Working with Big Data: Assembling Your Toolkit	8
Big Data for Sale: Data Marketplaces	10
Big Data: Courting Data Science Talent	11
First Steps on the Road to a Big Data Project	13
Examples of Big Data Projects	15
Recommended Reading: Books, Reports, Blogs, and Conferences	16

The first weeks of 2013 were barely out of the gate when one industry analyst was already predicting that Big Data would be *Time magazine's* 2013 "Person of the Year." Watching what Google, Amazon, and Facebook have done with Big Data is impressive, enviable.

But what about companies with smaller staffs and budgets? When does it make sense to start up a Big Data program? If your email marketing system isn't talking to your sales force automation system, and neither is synched up with your online purchase system, are you really ready to tackle a Big Data project?

The answer may surprise you as we examine Big Data's role in creating the next-generation digital experience.



DC Denison covered the technology scene for "The Boston Globe" for more than a decade, including serving as Technology Editor.

# What is Big Data?

At its most basic, Big Data is simply more data and more varieties of data than can be handled by a conventional database. But the term also refers to the many tools and techniques that have emerged to help users mine valuable information from these massive torrents of data. So it's not just the accumulation of information, it's the ability to analyze it for profit, insight, or both.

Many of these new tools have also lowered the cost of mining data. Open-source software, like Hadoop, and the availability of cloud services have dramatically lowered the price of a ticket to the Big Data Big Top. Big Data analysis has also been democratized because large data, and data and machine learning tools, are often available for free.

And fueling this revolution, of course, is the unprecedented amount of information that is now available for analysis and action: Every Facebook post or “like,” every beep of a supermarket scanner, every blink of a medical device can be sorted and analyzed, potentially yielding significant and valuable information. The ability to mine valuable information from these kinds of information torrents—looking for connections that are not obvious—is Big Data.

There is some fuzziness to the definition of Big Data. One common approach is to break it down into three Vs:

- **Volume.** Billions of computers, smartphones users, and objects are now operating and interacting with each other, generating exabytes (trust me, that's big) of data every day.
- **Variety.** Much of this data is “unstructured,” meaning that it doesn't notch neatly into a standard relational database. Unstructured data is a book review on Amazon, a comment on a blog, a video on YouTube, a podcast, a tweet.
- **Velocity.** A smartphone user's location data is changing constantly, so is the value of a portfolio held by a customer of an online financial service. These kinds of rapid updates present new challenges to information systems.

Recently **Mike Gualtieri**, a principal analyst with Forrester Research in Cambridge, MA (and the analyst who believes that Big Data will get *Time's* year-end honor), has come up with what he believes is a “more pragmatic” definition of Big Data, one that he claims is “an actionable, complete definition for IT and business professionals.”

Here's how Gualtieri defines it: “Big Data is the frontier of a firm's ability to store, process, and access all the data it needs to operate effectively, make decisions, reduce risks, and serve customers.”

Gualtieri says that the key to his definition can be summed up in the acronym SPA, which stands for: Store, Process, and Access. He advises clients to ask themselves these three questions to determine whether they are able to wring value from a Big Data project:

- **Store.** Can you capture and store the data?
- **Process.** Can you cleanse, enrich, and analyze the data?
- **Access.** Can you retrieve, search, integrate, and visualize the data?

Gualtieri's point is that businesses should define their Big Data projects not by the size or shape of the data, but by what they can accomplish—what they can do—with large, various, and fast-moving data.

The two definitions frame the issue. The three Vs describe what Big Data is in terms of size: how big is it, what it looks like. Gualtieri's definition describes what a company should expect to do with a Big Data project.



---

*“Big Data is the  
frontier of a firm's  
ability to store,  
process, and access  
all the data it needs  
to...serve customers.”*

— Mike Gualtieri,  
Forrester Research

---

# Should You Start a Big Data Project?

“If your organization stores multiple petabytes of data, if the information most critical to your business resides in forms other than rows and columns of numbers, or if answering your biggest question would involve a ‘mashup’ of several analytical reports, you’ve got a Big Data opportunity.”

That’s according to **Thomas H. Davenport**, a visiting professor at Harvard Business School who writes frequently about Big Data. Davenport advises companies to push beyond the buzzword to define their projects more precisely.

“Because the term is so imprecise,” he says, “organizations need to deconstruct it a bit in order to refine their strategies and signal to stakeholders what they are really interested in doing with these new types of data.”

For example, instead of saying, “We’re embarking on a Big Data initiative,” Davenport recommends that a company says, “We’re going to analyze video data at our ATMs and branches to better understand customer relationships.”

Similarly a health care organization can get more specific with its Big Data project by saying that it intends to “Combine electronic medical records and genomic data to create personalized treatment regimens for patients.”

Not only is this approach more precise, Davenport says, but it also avoids endless discussions about whether the data involved is big or small. (Few organizations, he points out, confess to working with “small data,” even though it’s a perfectly respectable activity).

**Bill Franks**, chief analytics officer for Teradata and a faculty member of the International Institute for Analytics, also advises companies not to be awed by the Big Data label.

“In many cases, Big Data is used for the exact same kind of analytics you’ve been doing for some time but with more data points from new data sources added to the mix.”

Franks, who is the author of the book *Taming the Big Data Tidal Wave*, (John Wiley & Sons, April 2012), points out that forward-looking companies are always struggling with new data types. In the late 1990’s and early 2000’s, for example, many organizations were struggling to use transactional data for broad analytics purposes. Now transaction data is “not much of a challenge,” he says.



---

*“Because the term is so imprecise, organizations need to deconstruct it a bit in order to refine their strategies...”*

— Tom Davenport,  
Harvard Business School

---

More recently, companies are getting used to working with online browsing history, a data type that was once considered daunting.

Big Data, according to Franks, is “simply a continuation of the struggle we’ve always had to incorporate ever-growing and ever more diverse data sources into analytics to enable better business decisions.”

That’s why the definition of Big Data, according to Mike Gualtieri, includes the word “frontier.” The push to incorporate ever larger and more various data is an essential part of a Big Data project.

To get the most from a Big Data project, experts say, you should start with a goal in mind. Do you want to measure the effectiveness of your marketing and advertising? How about incorporating the “voice of the consumer” in your product lifecycle decisions? Big Data can also be used to create brand new information products and services.

Be explicit about your business goals. That will shape your project, and increase the odds of a successful outcome. If you don’t have a goal, take a look at the “Examples of Big Data.” Maybe one of those project goals can be adapted to work for your company or organization.

Remember, too: Most often Big Data “aha moments” result from the intersections among a variety of data sources. Large collections of data tend to be stored in silos. Powerful new strategies and insights emerge when you cut across those vertical containers.

**Shawndra Hill**, who works with and teaches about Big Data in the Operations and Information Management Department at The Wharton School of the University of Pennsylvania, advises that a company, “should first understand the state of the art in data mining for their domain in order to identify the best benchmarks for their project and to see whether some existing solution is available to solve their problems.” The next step, according to Hill: “Calculate the expected gain from implementing the project in the best and worst cases of success and compare the estimates to the expected cost of taking on the project...No Big Data project should start just because it’s fashionable.”

The most successful Big Data projects are also “action-oriented,” with a strong internal push toward acting on the insights that emerge from the analysis. This is why consultants like Mike Gualtieri caution companies to avoid accepting Big Data projects that generate “lazy data.”

“If you have a data warehouse and you’re just producing reports, that’s not Big Data,” Gualtieri says. “You have to be able to use the information to create a competitive advantage in your markets.”



---

*“You have to be able to use the information to create a competitive advantage in your markets.”*

— Shawndra Hill,  
The Wharton School of the  
University of Pennsylvania

---

# Working with Big Data: Assembling Your Toolkit

The good news for those who want to tap into the power of Big Data: The rise of this data revolution has been powered by a number of prominent open-source and low-cost cloud computing projects, in addition to an explosion of commercial offerings. These projects are moving targets, constantly evolving (see “Recommended Reading: Books, Reports, Blogs, and Conferences” for ways to keep up), but a familiarity with the primary pieces of the Big Data puzzle will help you get oriented.

## HADOOP

The most important software in Big Data, and the one that sits at the white hot center of this revolution, is **Apache Hadoop**, an open-source project that runs on commodity Linux hardware.

Hadoop, which is named after a favorite stuffed elephant of the creator’s daughter, was developed at Yahoo, and was inspired initially by papers published by Google outlining its approach to handling an avalanche of data.

Hadoop implements a framework named Map/Reduce, where the application is divided into many small fragments of work, and it assigns that work to the nodes in a cluster. Hadoop also provides a distributed file system, HDFS, that spans all the nodes in a Hadoop cluster for data storage. HDFS links together the file systems on many local nodes to make them into one big file system.

Hadoop’s strength is that it can parallel process huge amounts of data across inexpensive, industry-standard servers that store and process the data. It can scale nearly without limits, which makes it uniquely suitable for working with ever-expanding sources of data.

Hadoop is also supplemented by an ecosystem of open-source Apache projects, such as Pig, Hive, and Zookeeper, that further extends the value of Hadoop and improves its usability.

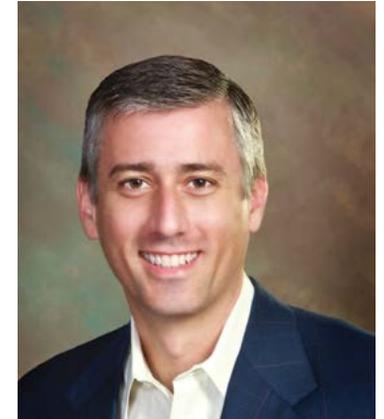
## SORTING OUT BIG DATA “SOLUTIONS”

Because Hadoop is open source, it has been incorporated into a wide variety of product offerings from the large, familiar enterprise vendors. Teradata has the **Aster Big Analytics Appliance**, EMC has **Greenplum**, IBM has **InfoSphere BigInsights**, Microsoft has its **Big Data Solution**, and Oracle offers a **Big Data Appliance**. There are also new Hadoop-based companies like **Cloudera**, and **Hortonworks**, and **MapR**.

Cloud computing has also come into the picture as an option for those considering a Big Data project. “Infrastructure as a Service” (or IaaS) providers enable users to buy time and install and configure their own software, like a Hadoop cluster. Budget-constrained companies can use these services to launch a Big Data project without having to invest in expensive hardware.

The next level up is cloud services that provide an application layer. Some of these “Platform as a Service” (PaaS) providers have already implemented Big Data solutions. Three major players are **Amazon Web Services**, **Google Cloud Services**, and **Microsoft Windows Azure**. Amazon Web Services and Microsoft’s Azure cross the boundaries between a service and platform, offering hybrid solutions. Google’s approach focuses on the application layer. California-based **Joyent** also offers hybrid products.

Because a lot of Big Data already lives in the cloud — such as data from social media and device sensors — cloud platforms are making more sense for hosting and analyzing Big Data. However, merging this data with what a company has on-premises will continue to be a challenge in the near term.



---

*“In many cases, Big Data is used for the exact same kind of analytics you’ve been doing for some time but with more data points...”*

— Bill Frank,  
Teradata

---

# Big Data for Sale: Data Marketplaces

Here's a break for companies hoping to put a Big Data project on the fast track: You don't have to generate all the Big Data you use. Often you can gain insights by purchasing another dataset and adding it to what you've been able to generate yourself.

For example, you can add weather data to your sales data to see if frigid temperatures are having a positive or negative effect on how your merchandise moves. And it's a buyer's market for data customers, with many marketplaces offering free datasets to compare with your own data streams.

Some of the most popular data marketplaces are **DataMarket**, **Factual**, **Microsoft Windows Azure Data Marketplace**, IBM, Google, Amazon, and Infochimps. At DataMarket, for example, you can search for datasets, gather the data, upload your own data, and compare. You can also output the results using DataMarket's chart and visualization templates.

Amazon offers **Public Data Sets** on AWS, which provides a centralized repository of free public data sets that can be integrated into AWS cloud-based applications. Google has **Public Data Explorer**, IBM's **ManyEyes** is geared toward visualization. Factual concentrates on places and products. You can mash-up your information with their data on local businesses, points of interest, restaurants, hotels, and consumer packaged goods. **Infochimps'** free datasets includes raw text of 4,771 erotica stories, 100,000+ official crossword words, and the birth and death rates of US teenagers, culled from the US Census.

Microsoft Windows Azure Data Marketplace, as the name implies, integrates data with its applications. Its data assets include economic indicators, telephone numbers, weather data, as well as regional datasets like crime statistics for England and Wales.

There are advantages to buying data from these marketplaces. For one thing, it's clean, which may be a welcome change from the messy data you've been trying to scrub. Many of the services also enable you to do your data crunching on their servers, freeing you from time-consuming and often complicated downloads. If you are already using a cloud-based data analytics solution from one of the providers, the process is even easier.

And you may be surprised by the variety of data that's available. "The wide availability of data continues to surprise me every day," said Shawndra Hill. "My colleagues and I have used publicly available data to predict drought in Ethiopia, the success of TV shows, what people will follow on Twitter, the success of advertising, and stock market trends. We have also worked on linking drugs to their side effects. In the past, these projects wouldn't be possible without partnerships with firms that allowed the use of their proprietary data."

# Big Data: Courting Data Science Talent

What skills do you need to implement a Big Data project? The field you want to explore is called “data science.”

“Data scientist” is a relatively new job title, but thousands already have it on their business cards (700 at Google alone). Yet because the field is so new—university programs are rare—Big Data professionals are hard to find. McKinsey & Co. predicts that by 2018, the U.S. could face a shortage of more than 1.5 million specialists needed to capture, store, manage, and analyze Big Data.

In Fall 2012, Thomas Davenport wrote a cover story for the *Harvard Business Review* that outlined strategies for staffing Big Data projects.

One approach they recommended: Grow your own. Recruit and develop Big Data talent in house, or look for achievers in any field with a strong data and computational focus and grow with them. Experimental physics and systems biology, for example, are two fields that could generate promising data scientists, according to Davenport and Patil.

But Davenport and Patil warn that the search won't be easy. What makes it particularly difficult, Davenport says, is that the best data scientists need a variety of technical, business, analytical, and relationship skills. According to Davenport, the best data scientists often have advanced computer science degrees, or advanced degrees in fields such as physics, biology, or social sciences that require a lot of computer work. In addition they have to be familiar with a wide variety of disciplines such as Hadoop, MapReduce and related tools, programming languages like Python, and disciplines like natural language processing.

Also, “Nothing beats experience,” adds Shawndra Hill. She says that the best data scientists have loved data for a long time and have gained an intuition about what can and can't be done. They also have a creative eye to think about how to use new data to solve old problems and old data to solve new problems.

---

*Davenport says, the goal is to find data talent that is “a hybrid of data hacker, analyst, communicator, and trusted adviser.” That combination, he admits, is “extremely powerful — and rare.”*

---

“Paths to data science usually start with an interest in solving hard problems,” Hill says. “The rest of the path is lined with exciting hard problems that have been solved successfully over time. The speed of computing makes so much more possible.”

In addition to these technical skills, data scientists also need the attributes previously necessary for analytical professionals, including mathematical and statistical skills, business acumen, and the ability to communicate effectively with customers, product managers, and decision makers.

The skills are so varied, Davenport reported, that some companies have decided to create data science teams that together embody this collection of skills. The yearly salary for data scientists, according to the online career site **Glassdoor**, ranges from \$80K to \$220K.

One encouraging sign for companies in search of expertise is that many of the hottest, most lavishly funded start-ups in the Big Data arena are working on products that mix analytics with Big Data, often in a cloud-based service. Ultimately these products could lighten the load for companies hoping to get a Big Data project off the ground. Until then, Davenport says, the goal is to find data talent that is “a hybrid of data hacker, analyst, communicator, and trusted adviser.” That combination, he admits, is “extremely powerful—and rare.”

# First Steps on the Road to a Big Data Project

“Start small with Big Data,” is the advice from author Bill Franks. Identify a few relatively simple analytics that won’t take much time or data to run. For example, an online retailer might start by identifying what products each customer viewed within just a few key categories so that the company can send a follow-up offer if they don’t purchase.

An organization that is entering the Big Data waters needs simple, intuitive examples to see what the data can do, Franks says, adding that this approach also yields results that are easy to test to see what type of lift the analytics provide.

Next, design a one-off test on some company data: a single month of data from one division for one set of products, for example. Franks cautions against attempting to analyze “all of the data all of the time” when first starting. That can muddy the water with too much data, and lead to high initial costs, a problem that plagues many Big Data initiatives. Instead, use only the data you need to perform the initial tests. At this point, Franks recommends, turn analytic professionals loose on the data. They can create test and control groups to whom they can send the follow-up offers, and then they can help analyze the results. During this process, they’ll also learn an awful lot about the data and how to use it.

Successful prototypes also make it far easier to get the support required for a larger, more comprehensive effort. Best of all, the full effort will now be less risky because the data is better understood and the value is already partially proven. It’s also worthwhile to learn early when the initial analytics aren’t as valuable as hoped. It tells you to focus your effort elsewhere before you’ve wasted many months and a lot of money.

“Pursuing Big Data with small, targeted steps can actually be the fastest, least expensive, and most effective way to go,” Franks says. “It enables an organization to prove there’s value in a major investment before making it, and to understand better how to make a Big Data program pay off for the long term.”

---

*To those who are on the fence, considering a Big Data project, Gualtieri has a simple piece of advice. “Don’t sit this out,” he urges. “This is real.”*

---

Whatever the size of your initial foray, experts advise to remember that it's a process, a loop. Don't expect fantastic insights the very first time you route two data streams into the same river. Often the benefits don't start to accrue until after you've run your tests through a few iterations.

Even then, because of the newness of the field, Big Data projects—even successful ones—can be frustrating. “We still have a ways to go to be able to combine evidence from different types of data sources—for example from text, social networks, and time series data,” says Shawndra Hill. “The methods have not caught up yet with the scale and complexities of today's Big Data.” She adds, “This is both exciting and scary. Exciting because there are a lot of new solutions to be generated, and scary because we are probably leaving a lot of value in databases, and that value may be harder to find as Big Data becomes even bigger data with even more complexity and noise.”

### Get Started

Analyst Mike Gualtieri likes to cite a Forrester study that predicts that by 2016, 1 billion people will have smartphones and tablets, “and that number will keep increasing,” he says. “The more technology people use, the more data they generate, and the more opportunity there is to provide personal experiences,” Gualtieri says. “The firms that make things personal will drive things in the future. The others will drop off.” To those who are on the fence, considering a Big Data project, Gualtieri has a simple piece of advice.

“Don't sit this out,” he urges. “This is real.”

# Examples of Big Data Projects

Here's another way to capture what a Big Data project could mean for your company or project: Study how others have applied the idea. Here are some real-world examples of Big Data in action:

- Consumer product companies and retail organizations are monitoring social media like Facebook and Twitter to get an unprecedented view into customer behavior, preferences, and product perception.
- Manufacturers are monitoring minute vibration data from their equipment, which changes slightly as it wears down, to predict the optimal time to replace or maintain. Replacing it too soon wastes money; replacing it too late triggers an expensive work stoppage
- Manufacturers are also monitoring social networks, but with a different goal than marketers: They are using it to detect aftermarket support issues before a warranty failure becomes publicly detrimental.
- Financial services organizations are using data mined from customer interactions to slice and dice their users into finely tuned segments. This enables these financial institutions to create increasingly relevant and sophisticated offers.
- Advertising and marketing agencies are tracking social media to understand responsiveness to campaigns, promotions, and other advertising mediums.
- Insurance companies are using Big Data analysis to see which home insurance applications can be immediately processed and which ones need a validating in-person visit from an agent.
- By embracing social media, retail organizations are engaging brand advocates, changing the perception of brand antagonists, and even enabling enthusiastic customers to sell their products.
- Hospitals are analyzing medical data and patient records to predict those patients that are likely to seek readmission within a few months of discharge. The hospital can then intervene in hopes of preventing another costly hospital stay.
- Web-based businesses are developing information products that combine data gathered from customers to offer more appealing recommendations and more successful coupon programs.
- The government is making data public at both the national, state, and city level for users to develop new applications that can generate public good.
- Sports teams are using data for tracking ticket sales and even for tracking team strategies

# Recommended Reading: Books, Reports, Blogs, and Conferences

## BOOKS

***Analytics at Work: Smarter Decisions. Better Results*** by Thomas H. Davenport, Jeanne G. Harris, Robert Morison, (Harvard Business Review Press; 2010)

***Taming the Big Data Tidal Wave*** by Bill Franks (John Wiley & Sons, 2012).

## REPORTS

***Data Scientist: The Sexiest Job of the 21st Century***, by Thomas H. Davenport and D.J. Patil (*Harvard Business Review*, Oct. 2012)  
HBR Reprint R1210D

***Big Data Now: Current Perspectives from O'Reilly Media***. Free download.

***Data Jujitsu: The Art of Turning Data into Product*** by D.J. Patil (O'Reilly Media, 2012).

## BIG DATA BLOGS/WEBSITES

Forrester Big Data Blog [[http://blogs.forrester.com/category/big\\_data](http://blogs.forrester.com/category/big_data)]

Greenplum.com [<http://www.greenplum.com/industry-buzz>]

Data Round Table [<http://www.dataroundtable.com/>]

Planet BigData [<http://planetbigdata.com/>]

Big on Data [<http://www.zdnet.com/blog/big-data/>]

IBM Smarter Computing Blog [<http://www.smartercomputingblog.com/category/big-data/>]

## CONFERENCES

Strata [<http://strataconf.com/strata2013/public/content/home>]

Structure:Data [<http://event.gigaom.com/structuredata/>]

Strata + Hadoop World [<http://strataconf.com/stratany2012/>]