CHAPTER 1

What Is Big Data and Why Is It Important?

Big Data is the next generation of data warehousing and business analytics and is poised to deliver top line revenues cost efficiently for enterprises. The greatest part about this phenomenon is the rapid pace of innovation and change; where we are today is not where we'll be in just two years and definitely not where we'll be in a decade.

Just think about all the great stories you will tell your grandchildren about the early days of the twenty-first century, when the Age of Big Data Analytics was in its infancy.

This new age didn't suddenly emerge. It's not an overnight phenomenon. It's been coming for a while. It has many deep roots and many branches. In fact, if you speak with most data industry veterans, Big Data has been around for decades for firms that have been handling tons of transactional data over the years—even dating back to the mainframe era. The reasons for this new age are varied and complex, so let's reduce them to a handful that will be easy to remember in case someone corners you at a cocktail party and demands a quick explanation of what's really going on. Here's our standard answer in three parts:

1. **Computing perfect storm.** Big Data analytics are the natural result of four major global trends: Moore's Law (which basically says that technology always gets cheaper), mobile computing (that smart phone or mobile tablet in your hand), social networking (Facebook, Foursquare, Pinterest, etc.), and cloud computing (you don't even have to own hardware or software anymore; you can rent or lease someone else's).
2. **Data perfect storm.** Volumes of transactional data have been around for decades for most big firms, but the flood gates have now opened with more *volume*, and the *velocity* and *variety*—the three Vs—of data that has arrived in unprecedented ways. This perfect storm of the three Vs makes it extremely complex and cumbersome with the current data management and analytics technology and practices.

3. **Convergence perfect storm.** Another perfect storm is happening, too. Traditional data management and analytics software and hardware technologies, open-source technology, and commodity hardware are merging to create new alternatives for IT and business executives to address Big Data analytics.

Let's make one thing clear. For some industry veterans, "Big Data" isn't new. There are companies that have dealt with billions of transactions for many years. For example, John Meister, group executive of Data Warehouse Technologies at MasterCard Worldwide, deals with a billion transactions on a strong holiday weekend. However, even the most seasoned IT veterans are awestruck by recent innovations that give their team the ability to leverage new technology and approaches, which enable us to affordably handle more data and take advantage of the variety of data that lives outside of the typical transactional world—such as unstructured data.

Paul Kent, vice president of Big Data at SAS, is an R&D professional who has developed big data crunching software for over two decades. At the SAS Global Forum 2012, Kent explained that the ability to store data in an affordable way has changed the game for his customers:

> People are able to store that much data now and more than they ever before. We have reached this tipping point where they don't have to make decisions about which half to keep or how much history to keep. It's now economically feasible to keep all of your history and all of your variables and go back later when you have a new question and start looking for an answer. That hadn't been practical up until just recently. Certainly the advances in blade technology and the idea that Google brought to market of you take lots and lots of small Intel servers and you gang them together and use their potential in aggregate. That is the super computer of the future.

Let's now introduce Misha Ghosh, who is known to be an innovator with several patents under his belt. Ghosh is currently an executive at MasterCard Advisors and before that he spent 11 years at Bank of America solving business issues by using data. Ghosh explains, "Aside from the changes in the actual hardware and software technology, there has also been a massive change in the actual evolution of data systems. I compare it to the stages of learning: dependent, independent, and interdependent."

Using Misha's analogy, let's breakdown the three pinnacle stages in the evolution of data systems:

- **Dependent** (Early Days). Data systems were fairly new and users didn't know quite know what they wanted. IT assumed that "Build it and they shall come."
- **Independent** (Recent Years). Users understood what an analytical platform was and worked together with IT to define the business needs and approach for deriving insights for their firm.
- **Interdependent** (Big Data Era). Interactional stage between various companies, creating more social collaboration beyond your firm's walls.

Moving from independent (Recent Years) to interdependent (Big Data Era) is sort of like comparing Starbucks to a hip independent neighborhood coffee shop with wizard baristas that can tell you when the next local environmental advisory council meet-up is taking place. Both shops have similar basic product ingredients, but the independent neighborhood coffee shop provides an approach and atmosphere that caters to social collaboration within a given community. The customers share their artwork and tips about the best picks at Saturday's farmers market as they stand by the giant corkboard with a sea of personal flyers with tear off tabs . . . "Web Designer Available for Hire, 555-1302."

One relevant example and Big Data parity to the coffee shop is the New York City data meet-ups with data scientists like Drew Conway, Jared Lander, and Jake Porway. These bright minds organize meet-ups after work at places like Columbia University and NYU to share their latest analytic application [including a review of their actual code] followed by a trip to the local pub for a few pints and more data chatter. Their use cases are a blend of Big Data corporate applications and other applications that actually turn their data skills into a helping hand for humanity.

For example, during the day Jared Lander helps a large healthcare organization solve big data problems related to patient data. By night, he is helping a disaster recovery organization with optimization analytics that help direct the correct supplies to areas where they are needed most. Does a village need bottled water or boats, rice or wheat, shelter or toilets? Follow up surveys six, 12, 18, and 24 months following the disaster help track the recovery and direct further relief efforts.

Another great example is Jake Porway, who decided to go full time to use Big Data to help humanity at DataKind, which is the company he co-founded with Craig Barowsky and Drew Conway. From weekend events to long-term projects, DataKind supports a data-driven social sector through services, tools, and educational resources to help with the entire data pipeline.

In the service of humanity, they were able to secure funding from several corporations and foundations such as EMC, O'Reilly Media, Pop Tech, National Geographic, and the Alfred P. Sloan Foundation. Porway described DataKind to us as a group of data superheroes:

> I love superheroes, because they're ordinary people who find themselves with extraordinary powers that they use to make the world a better place. As data and technology become more ubiquitous and the need for insights more pressing, ordinary data scientists are finding themselves with extraordinary powers. The world is changing and those who are stepping up to use data for the greater good have a real opportunity to change it for the better.

In summary, the Big Data world is being fueled with an abundance mentality; a rising tide lifts all boats. This new mentality is fueled by a gigantic global corkboard that includes data scientists, crowd sourcing, and opens source methodologies.

A Flood of Mythic "Start-Up" Proportions

Thanks to the three converging "perfect storms," those trends discussed in the previous section, the global economy now generates unprecedented quantities of data. People who compare the amount of data produced daily to a deluge of mythic proportions are entirely correct. This flood of data represents something we've never seen before. It's new, it's powerful, and yes, it's scary but extremely exciting.

> *The best way to predict the future is to create it!*
>
> —Peter F. Drucker

The influential writer and management consultant Drucker reminds us that the future is up to us to create. This is something that every entrepreneur takes to heart as they evangelize their start-up's big idea that they know will impact the world! This is also true with Big Data and the new technology and approaches that have arrived at our doorstep.

Over the past decade companies like Facebook, Google, LinkedIn, and eBay have created treasured firms that rely on the skills of new data scientists, who are breaking the traditional barriers by leveraging new technology and approaches to capture and analyze data that drives their business. Time is flying and we have to remember that these firms were once start-ups. In fact, most

of today's start-ups are applying similar Big Data methods and technologies while they're growing their businesses. The question is how.

This is why it is critical that organizations ensure that they have a mechanism to change with the times and not get caught up appeasing the ghost from data warehousing and business intelligence (BI) analytics of the past! At the end of the day, legacy data warehousing and BI analytics are not going away anytime soon. It's all about finding the right home for the new approaches and making them work for you!

According to a recent study by the McKinsey Global Institute, organizations capture trillions of bytes of information about their customers, suppliers, and operations through digital systems. Millions of networked sensors embedded in mobile phones, automobiles, and other products are continually sensing, creating, and communicating data. The result is a 40 percent projected annual growth in the volume of data generated. As the study notes, 15 out of 17 sectors in the U.S. economy already "have more data stored per company than the U.S. Library of Congress."[1] The Library of Congress itself has collected more than 235 terabytes of data. That's Big Data.

Big Data Is More Than Merely Big

What makes Big Data different from "regular" data? It really all depends on when you ask the question.

Edd Dumbill, founding chair of O'Reilly's Strata Conference and chair of the O'Reilly Open Source Convention, defines Big Data as "data that becomes large enough that it cannot be processed using conventional methods."

Here is how the McKinsey study defines Big Data:

> Big data refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze. This definition is intentionally subjective. . . . We assume that, as technology advances over time, the size of datasets that qualify as big data will also increase. Also note that the definition can vary by sector, depending on what kinds of software tools are commonly available and what sizes of datasets are common in a particular industry. With those caveats, big data in many sectors today will range from a few dozen terabytes to multiple petabytes (thousands of terabytes).[2]

Big Data isn't just a description of raw volume. "The real issue is usability," according to industry renowned blogger David Smith. From his perspective, big datasets aren't even the problem. The real challenge is identifying or developing most cost-effective and reliable methods for extracting value from

all the terabytes and petabytes of data now available. That's where Big Data analytics become necessary.

Comparing traditional analytics to Big Data analytics is like comparing a horse-drawn cart to a tractor–trailer rig. The differences in speed, scale, and complexity are tremendous.

Why Now?

On some level, we all understand that history has no narrative and no particular direction. But that doesn't stop us from inventing narratives and writing timelines complete with "important milestones." Keeping those thoughts in mind, Figure 1.1 shows a timeline of recent technology developments.

If you believe that it's possible to learn from past mistakes, then one mistake we certainly do not want to repeat is investing in new technologies that didn't fit into existing business frameworks. During the customer relationship management (CRM) era of the 1990s, many companies made substantial investments in customer-facing technologies that subsequently failed to deliver expected value. The reason for most of those failures was fairly straightforward: Management either forgot (or just didn't know) that big projects require a synchronized transformation of people, process, and technology. All three must be marching in step or the project is doomed.

We can avoid those kinds of mistakes if we keep our attention focused on the outcomes we want to achieve. The technology of Big Data is the easy part—the hard part is figuring out what you are going to do with the output generated by your Big Data analytics. As the ancient Greek philosophers said, "Action is character." It's what you do that counts. Putting it bluntly, make

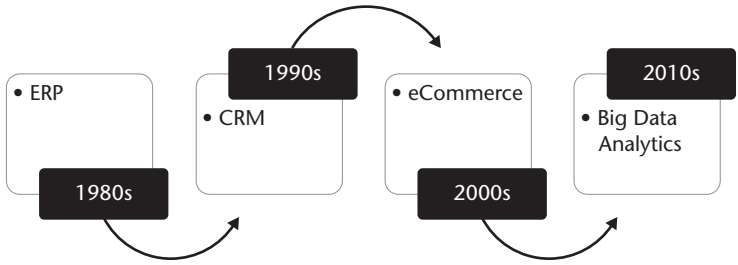

Figure 1.1 Timeline of Recent Technology Developments

sure that you have the people and process pieces ready before you commit to buying the technology.

A Convergence of Key Trends

Our friend, Steve Lucas, is the Global Executive Vice President and General Manager, SAP Database & Technology at SAP. He's an experienced player in the Big Data analytics space, and we're delighted that he agreed to share some of his insights with us. First of all, according to Lucas, it's important to remember that big companies have been collecting and storing large amounts of data for a long time. From his perspective, the difference between "Old Big Data" and "New Big Data" is accessibility. Here's a brief summary of our interview:

> Companies have always kept large amounts of information. But until recently, they stored most of that information on tape. While it's true that the amount of data in the world keeps growing, the real change has been in the ways that we access that data and use it to create value.
>
> Today, you have technologies like Hadoop, for example, that make it functionally practical to access a tremendous amount of data, and then extract value from it. The availability of lower-cost hardware makes it easier and more feasible to retrieve and process information, quickly and at lower costs than ever before.
>
> So it's the convergence of several trends—more data and less expensive, faster hardware—that's driving this transformation. Today, we've got raw speed at an affordable price. That cost/benefit has really been a game changer for us.
>
> That's first and foremost—raw horsepower. Next is the ability to do that real-time analysis on very complex sets of data and models, so it's not just let me look at my financials or let me look at marketing information. And finally, we now have the ability to find solutions for very complex problems in real time.

We asked Steve Lucas to offer some examples of scenarios in which the ability to analyze Big Data in real time is making an impact. Here's what he told us:

> A perfect example would be insurance companies. They need to know the answers to questions like this: As people age, what kinds of different services will they need from us?

In the past, the companies would have been forced to settle for general answers. Today, they can use their data to find answers that are more specific and significantly more useful. Here are some examples that Lucas shared with us from the insurance and retail industries:

> You don't have to guess. You can look at actual data, from real customers. You can extract and analyze every policy they've ever held. The answers to your questions are buried in this kind of massive mound of data—potentially petabytes worth of data if you consider all of your insurance customers across the lifespan of their policies. It's unbelievable how much information exists.
>
> But now you've got to go from the level of petabytes and terabytes down to the level of a byte. That's a very complex process. But today you can do it—you can compare one individual to all the other people in an age bracket and perform an analysis, in real time. That's pretty powerful stuff. Imagine if a customer service rep had access to that kind of information in real time. Think of all the opportunities and advantages there would be, for the company and for the customer.
>
> Here's another example: You go into a store to buy a pair of pants. You take the pants up to the cash register and the clerk asks you if you would like to save 10 percent off your purchase by signing up for the store's credit card.
>
> 99.9 percent of the time, you're going to say "no." But now let's imagine if the store could automatically look at all of my past purchases and see what other items I bought when I came in to buy a pair of pants—and then offer me 50 percent off a similar purchase? Now that would be relevant to me. The store isn't offering me another lame credit card—it's offering me something that I probably want, at an attractive price.

The two scenarios described by Lucas aren't fantasies. Yesterday, the cost of real-time data analysis was prohibitive. Today, real-time analytics have become affordable. As a result, market-leading companies are already using Big Data Analytics to improve sales revenue, increase profits, and do a better job of serving customers.

Before moving on, it's worth repeating that not all new Big Data technology is open source. For example, SAP successfully entered the Big Data market with SAP HANA, an in-memory database platform for real-time analytics and applications. Products like SAP HANA are reminders that suppliers of proprietary solutions, such as SAP, SAS, Oracle, IBM, and Teradata, are playing—and will obviously continue to play—significant roles in the evolution of Big Data analytics.

Relatively Speaking . . .

Big Data, as you might expect, is a relative term. Although many people define Big Data by volume, definitions of Big Data that are based on volume can be troublesome since some people define volume by the number of occurrences (in database terminology by the rows in a table or in analytics terminology known as the number of observations).

Some people define volume based on the number of interesting pieces of information for each occurrence (or in database terminology, the columns in a table or in analytics terminology the features or dimensions) and some people define volume by the combination of depth and width.

If you're a midmarket consumer packaged goods (CPG) company, you might consider 10 terabytes as Big Data. But if you're a multinational pharmaceutical corporation, then you would probably consider 500 terabytes as Big Data. If you're a three-letter government agency, anything less than a petabyte is considered small.

The industry has an evolving definition around Big Data that is currently defined by three dimensions:

1. Volume
2. Variety
3. Velocity

These are reasonable dimensions to quantify Big Data and take into account the typical measures around volume and variety plus introduce the velocity dimension, which is a key compounding factor.

Let's explore each of these dimensions further.

Data *volume* can be measured by the sheer quantity of transactions, events, or amount of history that creates the data volume, but the volume is often further exacerbated by the attributes, dimensions, or predictive variables. Typically, analytics have used smaller data sets called *samples* to create predictive models. Oftentimes, the business use case or predictive insight has been severely blunted since the data volume has purposely been limited due to storage or computational processing constraints. It's similar to seeing the iceberg that sits above the waterline but not seeing the huge iceberg that lies beneath the surface.

By removing the data volume constraint and using larger data sets, enterprises can discover subtle patterns that can lead to targeted actionable micro-decisions, or they can factor in more observations or variables into predictions that increase the accuracy of the predictive models. Additionally, by releasing the bonds on data, enterprises can look at data over a longer period of time to create more accurate forecasts that mirror real-world complexities of interrelated bits of information.

Data *variety* is the assortment of data. Traditionally data, especially operational data, is "structured" as it is put into a database based on the type of data (i.e., character, numeric, floating point, etc.). Over the past couple of decades, data has increasingly become "unstructured" as the sources of data have proliferated beyond operational applications.

Oftentimes, text, audio, video, image, geospatial, and Internet data (including click streams and log files) are considered *unstructured data*. However, since many of the sources of this data are programs the data is in actuality "semi-structured." Semi-structured data is often a combination of different types of data that has some pattern or structure that is not as strictly defined as structured data. For example, call center logs may contain *customer name + date of call + complaint* where the complaint information is unstructured and not easily synthesized into a data store.

Data *velocity* is about the speed at which data is created, accumulated, ingested, and processed. The increasing pace of the world has put demands on businesses to process information in real-time or with near real-time responses. This may mean that data is processed on the fly or while "streaming" by to make quick, real-time decisions or it may be that monthly batch processes are run interday to produce more timely decisions.

A Wider Variety of Data

The variety of data sources continues to increase. Traditionally, internally focused operational systems, such as ERP (enterprise resource planning) and CRM applications, were the major source of data used in analytic processing. However, in order to increase knowledge and awareness, the complexity of data sources that feed into the analytics processes is rapidly growing to include a wider variety of data sources such as:

- Internet data (i.e., clickstream, social media, social networking links)
- Primary research (i.e., surveys, experiments, observations)
- Secondary research (i.e., competitive and marketplace data, industry reports, consumer data, business data)
- Location data (i.e., mobile device data, geospatial data)
- Image data (i.e., video, satellite image, surveillance)
- Supply chain data (i.e., EDI, vendor catalogs and pricing, quality information)
- Device data (i.e., sensors, PLCs, RF devices, LIMs, telemetry)

The wide variety of data leads to complexities in ingesting the data into data storage. The variety of data also complicates the transformation (or the

changing of data into a form that can be used in analytics processing) and analytic computation of the processing of the data.

The Expanding Universe of Unstructured Data

We spoke with Misha Ghosh to get a "level set" on the relationship between structured data (the kind that is easy to define, store, and analyze) and unstructured data (the kind that tends to defy easy definition, takes up lots of storage capacity, and is typically more difficult to analyze).

Unstructured data is basically information that either does not have a predefined data model and/or does not fit well into a relational database. Unstructured information is typically text heavy, but may contain data such as dates, numbers, and facts as well. The term *semi-structured data* is used to describe structured data that doesn't fit into a formal structure of data models. However, semi-structured data does contain tags that separate semantic elements, which includes the capability to enforce hierarchies within the data.

At this point, it's fair to ask: If unstructured data is such a pain in the neck, why bother? Here's where Ghosh's insight is priceless. Our conversation with him was long and wide-ranging, but here are the main takeaways that we would like to share with you:

- The amount of data (all data, everywhere) is doubling every two years.
- Our world is becoming more transparent. We, in turn, are beginning to accept this as we become more comfortable with parting with data that we used to consider sacred and private.
- Most new data is unstructured. Specifically, unstructured data represents almost 95 percent of new data, while structured data represents only 5 percent.
- Unstructured data tends to grow exponentially, unlike structured data, which tends to grow in a more linear fashion.
- Unstructured data is vastly underutilized. Imagine huge deposits of oil or other natural resources that are just sitting there, waiting to be used. That's the current state of unstructured data as of today. Tomorrow will be a different story because there's a lot of money to be made for smart individuals and companies that can mine unstructured data successfully.

The implosion of data is happening as we begin to embrace more open and transparent societies. "Résumés used to be considered private information," says Ghosh. "Not anymore with the advent of LinkedIn." We have similar stories with Instagram and Flickr for pictures, Facebook for our circle of

friends, and Twitter for our personal thoughts (and what the penalty can be given the recent London Olympics, where a Greek athlete was sent home for violating strict guidelines on what athletes can say in social media).

"Even if you don't know how you are going to apply it today, unstructured data has value," Ghosh observes. "Smart companies are beginning to capture that value, or they are partnering with companies that can capture the value of unstructured data. For example, some companies use unstructured social data to monitor their own systems. How does that work? The idea is simple: If your customer-facing website goes down, you're going to hear about it really quickly if you're monitoring Twitter. Monitoring social media can also help you spot and fix embarrassing mistakes before they cost you serious money."

We know of one such "embarrassing mistake," when a large bank recently discovered that one of its ad campaigns included language that some people interpreted as hidden references to marijuana. The bank found out by monitoring social media.

Of course, not all unstructured data is useful. Lots of it is meaningless noise. Now is the time to begin developing systems that can distinguish between "%^*()334" and "your product just ate my carpet." In many ways, the challenges of Big Data and, in particular, unstructured data are not new. Distinguishing between signal and noise has been a challenge for time immemorial. The main difference today is that we are using digital technology to separate the wheat from the chafe. Companies like Klout have come up with influence scores that can be used to filter out pertinent data.

Talking to Misha Ghosh was a wake-up call. It's a reminder that now is the time to develop the experience that you will need later when the use of unstructured social data becomes commonplace and mainstream. In other words, learn as much as you can now, while there's still time to gain a competitive advantage, and before everyone else jumps on the bandwagon.

The growing demands for data volume, variety, and velocity have placed increasing demands on computing platforms and software technologies to handle the scale, complexity, and speed that enterprises now require to remain competitive in the global marketplace.

For a moment, let's forget about the definitions and technology underpinning Big Data analytics. Let's stop and ask the *big* question:

Is Big Data analytics worth the effort?

Yes, without a doubt Big Data analytics is worth the effort. It will be a competitive advantage, and it's likely to play a key role in sorting winners from losers in our ultracompetitive global economy.

Early validations of the business value are making their way into the public forum via leading technology research firms. For example, in December 2011, Nucleus Research concluded that analytics pays back $10.66 for every dollar spent, while Forrester produced a Total Economic Impact Report for IBM that concluded Epsilon realized a 222 percent ROI within 12 months

from a combination of capital expenditure (capex) and operational expenditure (opex) savings, productivity increase plus a revenue lift of $2.54 million.[3,4] In another example, Nucleus Research determined that Media Math achieved a 212 percent in five months with an annual revenue lift of $2.2 million.[5]

And, yes, there will be business and technology hurdles to clear. From a business perspective, you'll need to learn how to:

- Use Big Data analytics to drive value for your enterprise that aligns with your core competencies and creates a competitive advantage for your enterprise
- Capitalize on new technology capabilities *and* leverage your existing technology assets
- Enable the appropriate organizational change to move towards fact-based decisions, adoption of new technologies, and uniting people from multiple disciplines into a single multidisciplinary team
- Deliver faster and superior results by embracing and capitalizing on the ever-increasing rate of change that is occurring in the global market place

Unlike past eras in technology that were focused on driving down operational costs mostly through automation, the "Analytics Age" has the potential to drive elusive top-line revenue for enterprises. For those enterprises that become adept with Big Data analytics, they will simultaneously minimize operational costs while driving top-line revenues to net substantial profit margins for their enterprise.

Big Data analytics uses a wide variety of advanced analytics, as listed in Figure 1.2, to provide:

- **Deeper insights.** Rather than looking at segments, classifications, regions, groups, or other summary levels you'll have insights into *all* the individuals, *all* the products, *all* the parts, *all* the events, *all* the transactions, etc.
- **Broader insights.** The world is complex. Operating a business in a global, connected economy is very complex given constantly evolving and changing conditions. As humans, we simplify conditions so we can process events and understand what is happening. But our best-laid plans often go astray because of the estimating or approximating. Big Data analytics takes into account all the data, including new data sources, to understand the complex, evolving, and interrelated conditions to produce more accurate insights.
- **Frictionless actions.** Increased reliability and accuracy that will allow the deeper and broader insights to be automated into systematic actions.

SQL Analytics	Descriptive Analytics	Data Mining	Predictive Analytics	Simulation	Optimization
• Count • Mean • OLAP	• Univariate distribution • Central tendency • Dispersion	• Association rules • Clustering • Feature extraction	• Classification • Regression • Forecasting • Spatial • Machine learning • Text analytics	• Monte Carlo • Agent-based modeling • Discrete event modeling	• Linear optimization • Non-linear optimization

Business Intelligence

Advanced Analytics

Figure 1.2 Analytics Spectrum

Table 1.1 Big Data Business Models

Improve Operational Efficiencies	Increase Revenues	Achieve Competitive Differentiation
Reduce risks and costs	Sell to microtrends	Offer new services
Save time	Enable self service	Seize market share
Lower complexity	Improve customer experience	Incubate new ventures
Enable self service	Detect fraud	

Source: Brett Sheppard, "Putting Big Data to Work: Opportunities for Enterprises," *GigaOm Pro,* March 2011.

GigaOm, a leading technology industry research firm, uses a simple framework (see Table 1.1) to describe potential Big Data Business Models for enterprises seeking to exploit Big Data analytics.

The competitive strategies outlined in the GigaOm framework are enabled today via packaged or custom analytic applications (see Table 1.2) depending on the maturity of the competitive strategy in the marketplace.

While Big Data analytics may not be the "Final Frontier," it certainly represents an enormous opportunity for businesses to exploit their data assets to realize substantial bottom line results for their enterprise. The key to success for organizations seeking to take advantage of this opportunity is:

- Leverage all your current data and enrich it with new data sources
- Enforce data quality policies and leverage today's best technology and people to support the policies
- Relentlessly seek opportunities to imbue your enterprise with fact-based decision making
- Embed your analytic insights throughout your organization

Setting the Tone at the Top

When mounting an argument for or against something, it's always a good idea to bring out your best minds. It's safe to say that Dr. Usama Fayyad is one of the best minds in Big Data analytics. A world-renowned pioneer in the world of analytics, data mining, and corporate data strategy, he was formerly Yahoo!'s chief data officer and executive vice president, as well as founder of Yahoo!'s research organization. A serial entrepreneur who founded his first startup, Audience Science (formerly DigiMine) in 2000 after leaving

Table 1.2 Enabling Big Data Analytic Applications

	Improve Operational Efficiencies	Increase Revenues	Achieve Competitive Differentiation
Mature Analytic Applications	■ Supply chain optimization ■ Marketing campaign optimization	■ Algorithmic trading	■ In-house custom analytic applications
Maturing Analytic Applications	■ Portfolio optimization ■ Risk management ■ Next best offer	■ Ad targeting optimization ■ Yield optimization	■ In-house custom analytic applications
Emerging Analytic Applications	■ Chronic disease prediction and prevention	■ Customer churn prevention	■ Product design optimization ■ Design of experiments optimization ■ Brand ■ Product Market Targeting

Microsoft, he sold his second company, DMX Group, to Yahoo! in 2004 and remained on Yahoo!'s senior executive team until late 2008. Prior to starting up ChoozOn, he was founder and CEO of Open Insights, a data strategy and data mining consulting firm working with the largest online and mobile companies in the world.

Dr. Fayyad's professional experience also includes five years at Microsoft directing the data mining and exploration efforts and developing database algorithms for Microsoft's Server Division. Prior to Microsoft he was with NASA's Jet Propulsion Laboratory, where he did award-winning work on the automated exploration of massive scientific databases. He earned his Ph.D. in engineering from the University of Michigan, Ann Arbor, and holds advanced degrees in electrical and computer engineering and in mathematics. He is also active in academic communities and is a Fellow of both the Association for Computing Machinery and the Association of the Advancement of Artificial Intelligence; he is Chairman of the ACM SIGKDD.

We include all of that biographical detail to make it clear that what Dr. Fayyad says really matters. In particular, his insights into the differences between traditional methods for handling data and newer methods are quite useful.

From his perspective, one of the most significant differences is that with Big Data analytics, you aren't constrained by predefined sets of questions or queries. With traditional analytics, the universe of questions you can ask the database is extremely small. With Big Data analytics, that universe is vastly larger. You can define new variables "on the fly." This is a very different scenario from the traditional methodologies, in which your ability to ask questions was severely limited.

Why is the ability to define new variables so critically important? The answer is easy: In the real world, you don't always know what you're looking for. So you can't possibly know in advance which questions you'll need to ask to find a solution.

Dr. Fayyad uses the second Palomar Sky Survey, a comprehensive effort to map the heavens, as an analogy to explain the inherent problems of handling Big Data. The survey, also known as POSS II, generated a huge amount of data. Here's a summary of what Dr. Fayyad told us in a recent interview:

> Astronomers are really, really good at extracting structure from image data. They think of the Sky Survey as a way of collecting layers of resolution data about billions of stars and other objects, which is very similar to how businesses deal with their customers. You know very little about the majority of your customers, and the data you have is noisy, incomplete, and potentially inaccurate. It's the same with stars.
>
> When the astronomers need to take a deeper look, they use a much higher resolution telescope that has a much narrower field of view of the sky. You're looking at a very tiny proportion of the universe, but you're looking much deeper, which means that you get much higher resolution data about those objects in the sky. When you have higher resolution data, a lot of objects that were hardly recognizable in the main part of the survey become recognizable. You can see whether they are stars or galaxies or something else.

Now the challenge becomes using what you've learned from one narrow sliver of the sky to predict what you will find in larger sections of the sky. Initially, the astronomers were working with 50 or 60 variables for each object. That's way too many variables for the human mind to handle. Eventually the astronomers discovered that only eight dimensions are necessary to make accurate predictions. Dr. Fayyad explains how this impacts the level of accuracy:

> They struggled with this problem for 30 years until they found the right variables. Of course, nobody knew that they needed only eight and they needed the eight simultaneously. Meaning, if you dropped

> any one of the key attributes, it became very difficult to predict with better than 70 percent accuracy whether something was a star or a galaxy. But if you actually used all eight variables together, you could get up to the 90 to 95 percent level of accuracy level that's critical for drawing certain conclusions.

Nonscientific organizations, such as businesses and government agencies, face similar problems. Gathering data is often easier than figuring out how to use it. As the saying goes, "You don't know what you don't know." Are all of the variables important, or only a small subset? With Big Data analytics, you can get to the answer faster. Most of us won't have the luxury of working a problem for 30 years to find the optimal solution.

Notes

1. McKinsey Global Institute, "Big Data: The Next Frontier for Innovation, Competition, and Productivity," June 2011.
2. Ibid.
3. Nucleus Research, "Research Note: Analytics Pays Back $10.66 for Every Dollar Spent," Document L122, November 2011, http://nucleusresearch.com/research/notes-and-reports/analytics-pays-back-10-dot-66-for-every-dollar-spent/.
4. IBM Data Management and Forrester Consulting, "Total Economic Impact of IBM's Netezza Data Warehouse Appliance with Advanced Analytics," August 2011, http://ibmdatamag.com/2012/03/the-total-economic-impact-of-ibms-netezza-data-warehouse-appliance-with-advanced-analytics/.
5. Nucleus Research, ROI Case Study: IBM Smarter Commerce: Netezza MediaMath, Document L112, October 2011, www-01.ibm.com/software/success/cssdb.nsf/CS/JHUN-8N748A?OpenDocument&Site=default&cty=en_us.