CHAPTER 1.5

Managing the Risks and Rewards of Big Data

MATT QUINN
CHRIS TAYLOR
TIBCO

One of the biggest challenges of the term *big data* is deciding on a standard definition of what those words really mean. For many companies that have worked in an environment of large datasets, fast-moving information, and data that lack traditional structure, working in an environment of big data is just business as usual. In this chapter we will discuss how managing the growing challenge of data is not new for a regional healthcare organization in the Midwestern United States, a global logistics company, and a major American retailer. But for a majority of organizations, which have neither integrated data nor built a strategy around its use, the term *big data* itself is a way to express the sudden digitization of many things that have been with us forever but were not previously captured and stored as data. For most companies, big data represents a significant challenge to growth and competitive positioning. In some cases, it represents the survival of the business.

BIG DATA: RISKS AND REWARDS

Digitization itself is not new, but the maturation and availability of the Internet; the rapid growth of mobile computing; and, more recently, the addition of sensor data (data derived from devices that sense their environment) to the mix have all pushed the boundaries of how we think about data and its uses. The term *big data* represents the need for a new way of thinking but also implies new tools and new ways of managing data. Like many things, data can be used to do positive things for the world, but it can also be used to manipulate, embarrass, or repress. Data can be highly accurate and efficiently structured or unstructured, fragmented, and highly suspect. Data can also be managed well or carelessly. Big data, in its outsized properties, amplifies those effects. It is in those extremes that the risks and rewards of big data are decided.

THREE KEY BIG DATA TRENDS

As the world becomes more familiar with big data, three key trends that have a significant impact on those risks and rewards are emerging. First and foremost, *big data leverages previously untapped data sources.* Those sources are of several types. The first includes wearable devices that stream data about an individual and his or her surrounding environment on a moment-by-moment basis—such sensors include the applications on a smartphone that sense movement. The sensor in a runner's shoe is a very consumer-facing example, but business-facing sensors, which track all kinds of things, are proliferating very quickly. A pacemaker is a sensor that has been around a while (the newer models give feedback to healthcare workers).

The next type comprises connected sensors that instantly digitize and report what is happening in any moment and in any location. Examples of this type include the global positioning system (GPS) device that reports location back to a central computer or a user,

and devices in the soil of a farm that sense when and how much to irrigate. There are also sensors in trains, for example, that watch for signals that maintenance is necessary before a human could ever see them, such as brake heat, brake wear, movement in the rails, and so on. This new breed of sensors is coming into service and is connected to the Internet, making big data even bigger than human-generated information.

The third type of sensor provides constant reporting by machines that perform the work critical to our security, health, and lifestyle. Machines can be something as large as an aircraft or locomotive or they can be components of one of those things. Some of the most interesting of these sensors are the ones that measure the way an aircraft engine is performing mid-flight. Machines used to be purely mechanical but are increasingly computer controlled. Those computer controls mean not only that data are constantly being fed into machines but that they are also coming out of machines at a quickly increasing rate.

We have reached a point of information discovery that reveals correlation before causation, leaving researchers scratching their heads to find the underlying causes for correlations that data analysis clearly demonstrates. TIBCO's chief executive officer, Vivek Ranadive, is fond of saying that we have reached a point where we may know the "what" without knowing the "why."

The previously untapped information sources create a data ecosystem that can be modeled in a way that blends historical with in-the-moment information and is remarkably useful for anticipating the future. These models accurately predict such diverse outcomes as the spread of disease, the failure rate of aircraft components, and consumer behaviors. Big data's effectiveness is tightly coupled to an organization's ability to bring the right data together in the right moments that allow for the right response and outcome. Whatever we may know today, the continued discovery of previously untapped data sources will continue to change and improve our models, allowing us to better anticipate future events and to continue to increase our ability to affect desired outcomes.

The desire to affect outcomes brings about the second trend of big data: *the need for automation technologies.* Richard Hackathorn wrote about the value-time curve of information back in 2004 in "Real-Time to Real-Value," just as the world was becoming broadly and acutely aware of the explosion of data.[1] Hackathorn's curve describes the decreasing value of data over time as it passes through stages of use (Figure 1).

The challenge of the decreasing value of data over time has become even more meaningful in the age of big data. Today, the volume, velocity, and variety of data continue to push the curve down and to the right as organizations struggle to capture, analyze, and decide in a gradually more difficult environment. Added to this complexity is the increasing access to real-time data that leaves organizations in some industries attempting to reduce their response time to microseconds, understanding that this is a crucial part of being successful in their business.

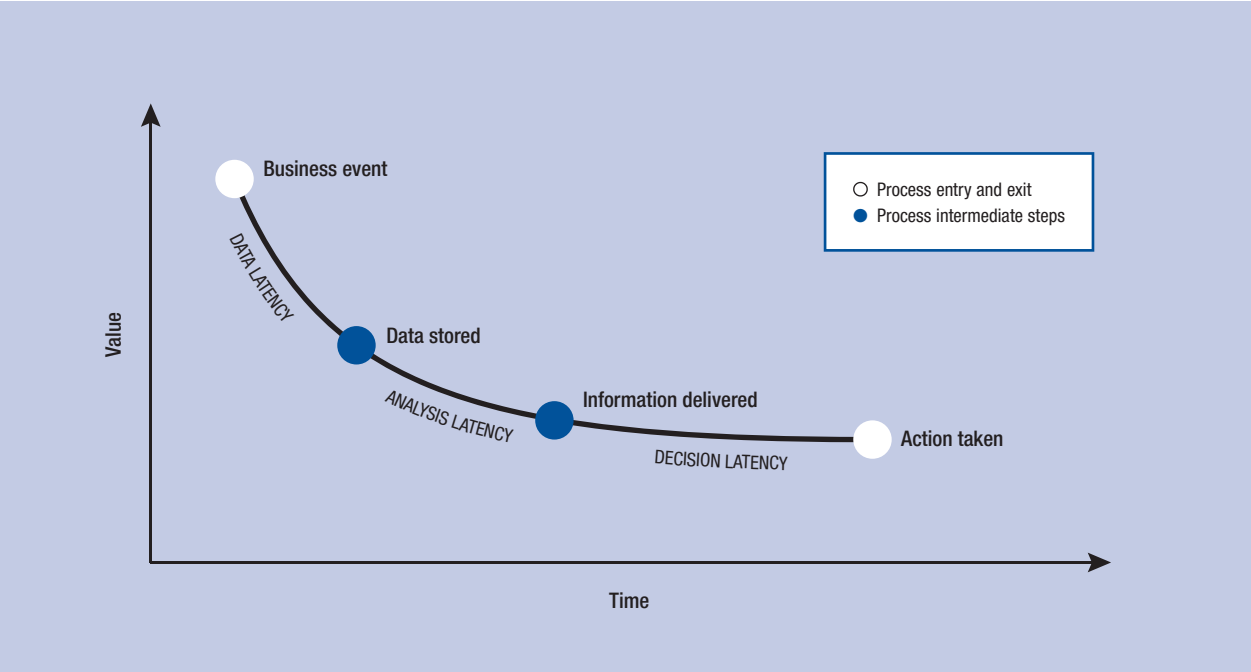
The value-time curve challenge makes big data management a function of creating automation wherever possible. Machines have always been humanity's friend in making work more efficient, and big data follows the same path. Big data's growth in each of its dimensions eliminates the ability for humans to intervene and reprogram processes in real time, opening the door for better and better tools that can manage data far more quickly and efficiently than a human can. Data exist in a moment, ready for decision and action, but there is a higher-level purpose for information. Data comprise the digital representation of events, or things that happen in patterns that occur over time, in conjunction with other events or in isolation, and even with things that may be expected but do not occur (such as when a patient fails to pick up a prescription after being discharged from a hospital, starting a likely string of events that will lead to readmission).

The idea of keeping track of what does *not* occur is a level of complexity higher than the old ways of waiting for data to arrive or change. Automation is especially well suited to the complexity of predicting, and then anticipating, events. In many organizations, automation is also a significant part of the actions that events precipitate.

The big data conversation often centers on the use of machines as the best resource for the storage and analytic processing of vast amounts of data, but this is only a piece of the story. Automation is increasingly a logical response to the need to find, filter, and correlate each piece of data as it flows over the enterprise so that decisions can be made—some through automation and some using a hybrid approach combining human and machine. Once decisions are reached, automation becomes the path for taking action in the shortest time frame possible before the value of data decays further.

The third trend being driven by big data is the *necessity for adaptable, less fragile systems.* For big data to leverage previously untapped sources of information, organizations need to quickly adapt to the opportunities and risks represented by these new sources. Automated systems that manage big data ecosystems cannot be developed around rigid schemas that require redevelopment for each new stream of information. Instead, systems need to absorb new information in an adaptable way that also adds value to existing data that have already been collected. Adaptable systems treat new sources of data coming constantly as the means to improve analytical models, create better decisions, and drive more appropriate actions.

Figure 1: The value-time curve



Source: Hackathorn 2004.

RESOLVING TWO PRIMARY CHALLENGES OF BIG DATA

Most organizations need to overcome two primary challenges before becoming productive with big data. The first is the need for powerful visualization that allows the business to explore data to find questions worth answering. This stands the traditional business intelligence model on its head, as the pre–big data model began with the business asking a question and ended with information technology structuring data to answer those questions in a very repeatable way, typically as dashboards. Visualization instead begins with capturing all data available so that multi-structured and iterative discovery can take place that reveals information with or without having the right question. Visualization lets the data speak for themselves.

Humans are extremely well suited to visual analysis. Our brains are wired to very rapidly assimilate what we see and spot patterns. Using our eyes, we can spot a trend or an outlier in a fraction of a second, far faster than we can by sifting through numbers on a screen. If a picture is worth a thousand words, visualization is worth petabytes, terabytes, and more of raw data. Visualized data and the human mind make for a highly efficient combination. Most importantly, visualized data have the effect of engaging the non-technical but business-savvy human in the iterative process of discovering exploitable insight. This lessens the organization's reliance on technical resources and, specifically, on data scientists.

The second hurdle that organizations face is the need to manage ever-larger amounts of data. Systems scoped for today's needs quickly become insufficient when the data are increasing in size, speed, and complexity. Unfortunately, when people talk about "big data" they often use the term to compartmentalize it and give it boundaries. This is a natural reaction and harkens to the beginning of computerization when data were processed as batches of transactions that represented a finite amount of information. Thinking of big data in those terms fails to take into account all of the data being created everywhere, every day. This compartmentalized view can also deprecate data that may not appear useful or valuable or may be difficult to process. At a point in the future, organizations will very likely look back and wish they had considered all data when deciding what to store. When we consider data without specific boundaries, we can focus our efforts on linking data together and analyzing them more broadly. We will probably find the data have value for a wider range of people in the organization than originally anticipated.

When we consider all data, we can see the value of discovering the connectivity of data. This brings into consideration different data types that are used to adorn our original data and make them more valuable as a source of visual, predictive, and operational analytics. Why does that matter? We have grown accustomed to having instantaneous answers to our questions. As data grow, they have the very real likelihood of slowing down how decisions are made. Nonlinear growth taxes our systems and creates the scenario that every day we get bogged down more as untapped data sources become newly available, our clever automations become less effective, and our systems seem less adaptable than before. An all-data approach allows the organization to see today's information as the best we have in the moment, knowing that we will continue to layer on more

data—not with the goal of having a larger dataset, but instead with the goal of using all of the data available to gain the best outcome. Rather than slowing down the results, using all available data takes into account data linkages and permits a broad analysis that allows the most organizational clients to constantly arrive at the best possible outcome.

Enabling the organization with visualization and the constantly additive benefits of all data allows experts to be able to explore data to find their value. For a retailer, that means being able to explore diverse data that include historical visits to the website as well as transactions completed or shopping carts abandoned; with the addition of geographical information from a mobile society, the retailer has an ability to understand the ambient circumstances at the time decisions are being made.

ENSURING THAT HUMANS STAY IN THE LOOP

For exactly this reason we need to take a very careful approach to how big data is being used and apply the right level of oversight. There are two specific reasons for having an appropriate governance model, each tackling the problem from the opposite perspective. The first is a need to ensure that data are not being used in a way that goes against the organization's best interests. Such unfortunate (even inappropriate) uses can be the result of rogue individuals with no checks and balances on their access and actions, or it can be the result of individuals acting with the best intentions but incurring unintended consequences that go against the goals of the organization. Data are very powerful, and organizations need to ensure that information is being collected, stored, analyzed, and acted upon in ways that can be audited and that raises alarms when necessary.

The second need for governance is demonstrated by the danger of having machines talking to machines without a human supervising the conversation. Systems need to leave an aperture for control by humans to avoid the problems of passive neglect or runaway processing. Finding the right balance is the challenge, and it involves looking at the value of the decisions being reached and the risk associated with the decision. There is a broad spectrum of judgments that covers small, incremental decisions that have moderate impact on an overall risk profile versus large, occasional decisions that can have enormous impact. Machines are exceptionally good at monitoring and executing detail, but the need for humans to focus on the macro decisions is significant. Consider the car analogy: a human cannot be involved in every firing of every cylinder. The human has absolute responsibility, however, for the speed of the car under the current conditions, monitoring the engine temperature, and a host of other variables.

STRIKING A PRIVACY BALANCE

We have watched the sharing of personal data increase year after year since people first connected across the Internet. Many of the risks and rewards of big data are coupled tightly to the use of all of those data. On the reward side, data can be used to create far better customer service by knowing the customers' needs and histories. They can be used to create more personalized offers based on customers' preferences and their loyalty to a brand. From this perspective, data can be used to engage the customer and to create a better relationship that serves everyone's needs. Healthcare-related personal information improves treatment and saves lives both at the individual level and in aggregate, as clinical trials of sample patients give way to all data about every patient.

Personalization and healthcare offer two standout opportunities for big data to reward us. At the same time, big data comes with privacy concerns that are not simply related to technology but are also about very human things such as privacy, all-knowing "creepiness," and personal security. Given enough personal data, information can be correlated that can be both unsettling and unwanted. Today's public, legislative, and legal sentiments may not be tomorrow's, and these attitudes tend to diverge by government and region of the world. What is standard practice in terms of collecting personal information in the United States is frowned upon in many parts of Europe. Managing the "Facebook Effect," where people willingly share ever-increasing amounts of personal information, is a challenge for individuals and governments as well as for the software companies that sit in the middle, confronted with inconsistent norms and laws across different locations in the world.

Privacy paradigms are in constant flux, but the need for a consistent approach to meet privacy expectations never changes. Protecting privacy has, at its roots, the need to protect data both at a discrete level and, maybe even more importantly, at an aggregate level. Learning a great deal about a person by combining factors that may seem harmless at a discrete level but, when taken together, may give away information that the person would not want generally known is one such example. This could happen, for instance, by combining someone's Facebook status with the location where he or she logged in to pay an electric bill with the home zip code; this could target wealthy people by knowing that they are not at home, making them vulnerable to burglaries. Each discrete piece of information is not meaningful, but in the aggregate can make someone a potential victim.

Systems exist that can manage the access, movement, and dissemination of data, but in our haste to build out the largest datasets and the maximum computational power, the need to put the right controls in place has been consistently overlooked. Some of this has been naiveté, and some has been a deliberate

stretching of the boundaries of individual expectations. Throughout the evolution of big data, the capability to govern data appropriately has existed, but unless organizations make the choice themselves or are pushed by legal or public pressure, the protection of personal privacy remains a low priority.

SHOWING BIG DATA'S SOPHISTICATED SYSTEMS

Gaining benefits from big data while mitigating risks is entirely a matter of data systems sophistication. This section will explore three examples that demonstrate the successful use of big data.

The first example of that sophistication is on display at a major network of hospitals in the Midwest to address the problem of sepsis—the systemic infection of the body—which is a constant threat to hospitalized patients. Sepsis is usually acquired in the healthcare facility; it is not the reason a patient arrives. Instead, sepsis appears somewhere between a patient's travel between the emergency room, the laboratory, the radiology department, and any other department where treatment is given. If not treated immediately, sepsis usually results in the death of the patient.

This healthcare company realized that, in order to tackle the sepsis problem, they had to create a sophisticated system that could follow a patient throughout his or her stay. The system needed to track patient data despite that patient's location within the hospital and despite the siloed information technology systems that are all too common in healthcare. Most of all, the system needed to bring data together in a way that allows high-speed correlation, based on prior analysis of sepsis data, so that medical staff can be alerted within life-saving time frames. This company's sophisticated system was successful at significantly shortening time frames for response to sepsis and significantly decreased the mortality rate in their facilities. They were successful enough, in fact, to allow their system to be turned into a Software-as-a-Service and contracted to other facilities.[2]

The second example is one of logistics. Like healthcare, logistics is an age-old practice undergoing big data transformation. It has become far more complicated in recent years because of the explosion of data that connect the customer's customer and the supplier's supplier. We are able to know significantly more in the form of digital data that not only allow the prediction of outcomes but that also allow us to make operational decisions at any point along the supply chain. For a global package delivery company, knowing their business means being able to access all available data to monitor not just the arrival and departure of aircraft but also the aircraft altimeter and attitude in order to provide additional layers of data that provide better insight on the nuanced status of the flight.[3] In a similar fashion, today's complex contracts encompass the global movement of pharmaceuticals and other sensitive cargos that require constantly monitoring all data. A global logistics company must monitor discrete data such as package temperature, location, and time to delivery that continually describe a shipment's ambient conditions; furthermore, these data must be available alongside expiration data and acceptable data ranges.

Those aggregate data form the basis for ensuring non-stop compliance to local and international standards for moving items that require special handling. Those same data ensure that contract terms are being respected and provide the basis for improving profitability while decreasing waste and inefficiency within a contracted service. It is a gift that keeps on giving, as detailed historical shipment data allow better pricing of potential new contracts, making the logistics carrier more competitive and reducing the risk of negotiating and accepting poor contracts. Without the ability to manage all relevant data, logistics companies and their customers would be unable to effectively move cargoes that bring enormous benefits to all parts of the planet.

The third example is seen in retail markets. In retail, the management of big data supports a brand's ability to predict the best product offering and to establish effective marketing and loyalty programs. It also supports better ways to sell and greatly improves customer service execution.[4] Big data offers an enormous reward to retail because successful selling is ultimately about having an excellent understanding of customers and the circumstances in which they buy. Even more importantly, successful retail is about creating the circumstances that *turn a customer into a fan.* A fan feels a personal connection to the brand and is much more likely to be an advocate. From a revenue perspective, a fan has a much greater total lifetime value.

But creating a fan is not a simple exercise in better customer service. Predictive analytics, heavily dependent on powerful visualization, form the basis for knowing the best moments and the best ways to engage with the customer. Understanding the past is key to predicting the future, and visualization reveals the meaningful patterns in data that tell us what happened under a host of variables in the past. Visual analytics tell the retailer what can be anticipated in today's real-time situations and set the stage for blending information streaming constantly from the website, store, and logistical systems, along with data coming from mobile devices. That information is vitally important to knowing not only how to provide information and offers to help a customer through a purchase, but also how to best serve a customer's needs after products have been purchased. The brand that knows its customers using this approach is leaps and bounds ahead of the one that lacks these capabilities.

Although the rewards are clear, a risk remains in gaining the customer's favor while requiring access to so much personal information. Loyalty programs are the

ideal way to gain that access and avoid the creepiness factor. Focused customer loyalty management elicits the customer's permission through a system of rewards and exclusive offers that provides benefit back to the customer, mitigating the risk of a brand being perceived as stalking the customer or invading their privacy.

ENSURING THE BENEFITS, MITIGATING THE RISKS

Managing the three key trends of leveraging previously untapped data sources, using automation wherever possible, and creating less fragile data systems are crucial parts of ensuring the benefits of big data while mitigating its risks. Accomplishing these three objectives requires successfully meeting big data's two main challenges: the need to visualize by using analytics tools and the need to systematically discover, capture, govern, and secure ever-larger amounts of data.

Big data has a remarkable ability to change the world. Its benefits need to be considered as a function of how well its risks are managed. Truly expert handling of big data brings the reward of being able to react to world-changing events, both big and small, at an unprecedented rate and scope. Epidemics can be tracked and miracle drugs developed, but at the same time, there is a need to ensure that humans are not cut out of the loop. Organizations need to carefully plan for the right level of oversight that gives an aperture of control to humans—after all, big data should be working for the benefit of humans, not the other way around.

Organizations that manage big data have an obligation to monitor security device, server, and application logs, all of which generate machine data that provide insight into how, when, and why machines are communicating with other machines. Monitoring the activities of machines allows organizations to watch for patterns and avoid runaway transactions or manipulation that can lead to fraud and other unintended results. Server logs also provide indications of who accessed data and how these data were used, affording critical oversight into potential illegal or unethical access and use of data. Machine data are monitored by healthcare organizations to show compliance with Health Insurance Portability and Accountability Act (HIPAA) standards, banks to prevent credit card fraud, and governments and corporations to watch for and prevent data loss.

Today's public, legislative, and legal sentiments may not be tomorrow's; these attitudes will continue to diverge by government and region. Governments and other organizations need to balance the Facebook Effect, which entails the deliberate sharing of more and more personal information, with the requirements of security and what the marketplace can use for better customer service and marketing. Organizations, both public and private, need to proactively take steps to prevent privacy intrusion whether the public demands such measures or not. European governments provide an example with the "right to be forgotten" for minors across the European Union. Those steps may include obtaining approval, either by asking permission or by gaining permission in exchange for tangible benefits for the collection and use of personal data—a common technique used by customer loyalty programs. Organizations should also consider the use of anonymization techniques to mask personal identities where that is the appropriate path.

Organizations, both public and private, must balance the risks and rewards of big data—especially as big data moves from low impact "experiments" to driving real-time operations and decision-making. Although social acceptance of what data can and will be shared is changing and evolving, its impact on privacy and personal security and the introduction of the creepiness factor are all things to consider. Big data is a fast-moving technology space that will affect all aspects of our lives. Transparency about what, how, and why data will be used will become more important as organizations seek to provide better services and products at both the government and private levels. Taken together, the trends and challenges will shape the path forward for organizations that wish to be deliberate and wise about their use of big data.

NOTES

1 Hackathorne 2004.

2 The website for the service is http://mercytelehealth.com/services/safe-watch/.

3 Confidential client example.

4 Confidential client example.

REFERENCES

Hackathorne, R. 2004. "The BI Watch: Real-Time to Real-Value." *DM Review,* January (2004). Available at http://www.bolder.com/pubs/DMR200401-Real-Time%20to%20Real-Value.pdf.

Mercy Services. Telehealth Services, Safe Watch. Available at http://mercytelehealth.com/services/safe-watch/.