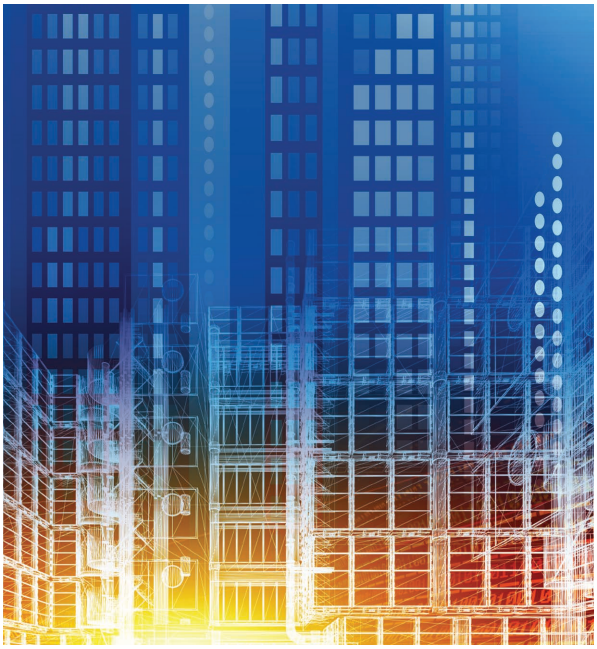

Corporate Governance of Big Data: Perspectives on Value, Risk, and Cost

Paul P. Tallon, *Loyola University Maryland*

Finding data governance practices that maintain a balance between value creation and risk exposure is the new organizational imperative for unlocking competitive advantage and maximizing value from the application of big data.

Organizations continue to experience exponential growth in the amounts of data they capture and retain within their datacenters. Companies such as Intel, Google, and Wal-Mart currently manage multiple petabytes of data that individually are hundreds of times larger than the content of the US Library of Congress.

A recent study by the Center for Large-Scale Data Systems at the University of California, San Diego, reported that the quantity of data in corporate datacenters continues to grow, on average, by 40 percent annually.[1] In some sectors—most notably healthcare, pharmaceuticals, energy, telecommunications, and transportation—managers report datacenter growth in excess of 100 percent per annum. Although exponential data growth is the new norm, remarkably, many organizations have failed to answer two important questions: Why does big data exist, and what is its true value?

The assumption guiding data capture and retention has always been that big data's value must exceed its cost. However, this belief is rarely, if ever, tested. Without a clear sense of value and how it can shift over time, organizations are liable to make mistakes. This could lead to extreme levels of technical, economic, or reputational risk if organizations underinvest in storage technology for highly valuable data—for example, storing clinical trial data for a new blockbuster drug on an unreliable storage device would be relatively inexpensive but risky. Similarly, organizations could waste resources by overinvesting in storage technology when the value of their data is low—for example, needlessly replicating data or using the most expensive and resilient storage device when a less expensive system would suffice.

The challenge facing organizations is to develop governance mechanisms—policies and structures—that strike a balance between risk and reward in the face of growing quantities of data and innovation that delivers better, faster, and cheaper storage technology. These policies and structures should protect data from errant factors that could destroy or limit its value, but governance should not be so restrictive and intransigent that it prevents or impedes organizations from unlocking value from their data.

Our research at Loyola University Maryland suggests that data governance is a reflection of how organizations value their diverse and expanding portfolio of data assets as well as their desire to invest in storage technologies to protect those assets from various risk factors. Because a one-size-fits-all approach is unlikely to succeed,

 Published by the IEEE Computer Society 0018-9162/13/$31.00 © 2013 IEEE

organizations should choose from a broad list of possible data governance practices modified for their particular industry and circumstances.

VALUING INFORMATION ASSETS

In the course of certifying annual financial statements, auditors are expected to confirm that valuable data exists and is adequately protected, just like any other corporate asset. It is worth noting, however, that financial reporting standards preclude valuing data assets for balance sheet purposes except for mergers or acquisitions wherein a company might be deliberately acquired to gain control of those information assets.

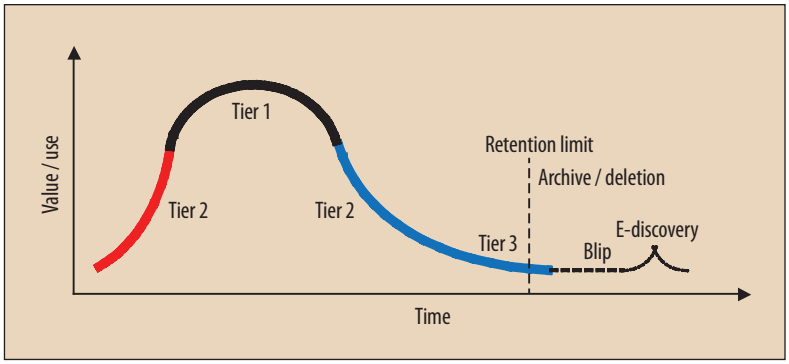

Figure 1. Information life-cycle curve. The value of information rises and falls over the course of time commensurate with its usefulness in organizational decision making.

When information assets are omitted from the balance sheet, they are treated as having zero value even in the most data-rich organizations. In the absence of an express or justifiable requirement for valuing data, organizations have failed to create methodologies that accurately measure the value of information or that track changes in value over time. Consequently, they struggle to know where their portfolio of information assets is on the information life-cycle curve in Figure 1 or when it is appropriate to migrate data between storage tiers consistent with changes in data value.

A practical application of the information life cycle requires an understanding of the length of time over which value rises and falls. However, value is often in the eye of the beholder, leading some users to retain data long after others consider it to be worthless. The expected life or duration of an information asset can vary widely. Stock prices can exist for a fraction of a second before their value decreases, whereas hospitals are expected to retain electronic medical records for the patient's life, which, by current standards, is roughly 75 to 80 years. Designing data governance practices to manage data over such extreme periods is especially complicated.

Although organizations lack a clear understanding of the value of their data, other insights are still possible. For example, managers often describe in detail how much value is lost—in the form of profits, revenues, or incremental data recovery costs—when critical data is lost or inaccessible for specified periods.[2] Datacenter managers put downtime costs due to storage issues at $42,000 per hour, but the range of estimates can be as high as $500,000 per hour in the case of airline reservation systems or highly used websites.[3]

By knowing how much value is at risk from adverse events such as hardware failures or accidental data destruction, datacenter managers can make strategic decisions to prevent repeat incidents that destroy value. For example, if a black swan event costing $1 million in lost revenues, profits, and damaged reputation could have been prevented with more fault-tolerant technologies or more advanced technical training, an additional $1 million storage investment would be fully justified. However, it would not make sense to spend $5 million to prevent a $1 million loss, nor would it be rational to spend amounts that still leave the organization exposed to preventable losses.

If we take the view that storage technologies and the other costs of managing big data are analogous to an insurance policy, then storage investments are the premiums required to manage risk. Lower insurance policy premiums can allow risks to increase, whereas higher premiums can reduce risk levels; policy owners choose an acceptable deductible and adjust spending levels accordingly. When storage investment decisions are seen in this way, spending can be specified for each type of data, but the critical assumption is that managers can rationally estimate how much value is potentially at risk from the loss or inaccessibility of data for each time period.

Unfortunately, not all data types or applications lend themselves to this type of value-at-risk analysis. Organizations readily capture and retain data that might yield little or no value in the near term but that could rise in value at some point in the future.

In some cases, anticipation of value can be sufficient to justify data retention for extended periods of time even when doing so poses a risk to personal privacy. Retention can be justified if the costs of storing the data and safeguarding its privacy are less than the opportunity costs of not having the data available for future decision making or analysis.

As such, managers might need to continually assess whether to retain data they believe could be valuable in the future—balancing the opportunity cost of not keeping the data against its retention costs. When the opportunity cost falls below the expected costs of retention, it might signal the need to begin removing or destroying data or to consider less expensive long-term storage technology.

Our research suggests that the risk of data loss is particularly a concern to senior executives in the case of

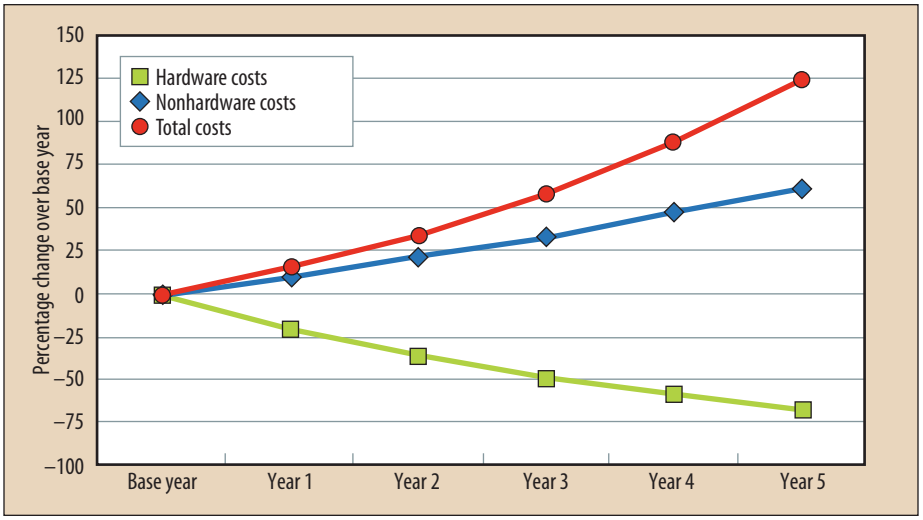

Figure 2. Big data cost simulation. Even with the most conservative cost and data growth estimates, total costs will more than double over a five-year period.

e-discovery, wherein organizations are asked to deliver data such as email or other correspondence to opposing counsel as part of a legal dispute. These concerns affect whether data should be retained beyond the end of its useful life if there is a possibility that it could be requested at a later date.

As Figure 1 shows, e-discovery blips might occur beyond the point where data can reasonably be archived or deleted. Value might temporarily increase not because the data helps improve organizational decision making but because failure to produce it could trigger legal damages. Such was the case in 2005 when damages totaling $1.45 billion were awarded against Morgan Stanley for failing to deliver specific email records as the court requested. Although this award was reversed on appeal, legal counsel and CIOs are reluctant, on the basis of what happened in that particular case, to discard data that might need to be produced in court. In some organizations, this concern has prompted corporate legal counsel to enforce perpetual retention of all data.

UNDERSTANDING COSTS

The belief that storage is cheap and its cost is likely to decline further is true in terms of hardware prices, but it ignores a variety of other datacenter operating expenses that comprise the total cost of ownership (TCO). Other costs such as energy, maintenance, and software increase in direct proportion to the growth of big data.

Estimates vary, but research confirms that TCO is approximately five to seven times higher than hardware acquisition costs.[4] Users and application owners who play a critical role in determining whether to collect and retain data rarely see these non-hardware costs and tend to believe they will be small relative to other IT expense items or, due to innovation, should be comfortably offset by declining hardware prices.

If users feel that the cost of retaining data is zero or close to zero, they will not be motivated to carefully decide what data to keep and for how long. Instead, there is an urge to keep as much data as possible on the basis that it might be useful in the future: "just in case" logic becomes the retention norm.

Moreover, users are more likely to add to the glut of big data when they are not responsible for covering the costs of managing it. Without some form of chargeback, the actual cost of managing data will be forever hidden from users. The tragedy of the commons argument is that if something has zero cost, it will be misused to the eventual detriment of all users. Storage is not quite at that point, but the rapid growth in big data is exacerbating the problem.

Some organizations are nearing the point where storage costs are so significant that they compromise strategic IT investing. Figure 2 shows a five-year projection for hardware costs, nonhardware costs, and combined costs for a hypothetical organization experiencing a 50 percent annual data growth rate. Because organizations do not replace their IT hardware annually to take advantage of improved price performance, we can apply a conservative 20 percent decline in hardware costs alongside a 10 percent rise in nonhardware costs for increased salaries, licensing, and operating expenses.

IT budgets are already under increased pressure. If organizations cannot contain their storage spending, there will be little alternative but to transfer resources from other IT projects. Research shows that 47 percent of IT budgets are assigned to maintain IT infrastructure (including storage hardware and networking), 40 percent to information and transaction processing, and 13 percent to strategic IT investments.[5]

The costs of managing big data and the evolution of data analytics programs are likely to cause organizations to spend significantly more on IT infrastructure and information processing, so flat or declining budgets could undermine strategic IT projects.[6] Storage was seen as tactical when data growth was minimal in the closing years of the 20th century and hardware and software were the primary sources of competitive advantage. But as data becomes increasingly strategic, and hardware and software become more commoditized and tactical,

CIOs may need to separate investments in managing the growth of big data from other budget line items.

The simplest form of cost control within a big data environment involves the use of storage tiers that span different areas of the information life cycle depicted in Figure 1. Organizations can use highly resilient, reliable, and accessible (but expensive) storage technologies for their most valuable data and less reliable or accessible (but cheaper) technologies for their relatively less valuable data and archives.

Data can migrate from lower tiers to higher tiers as value rises. A higher tier offers lower risk that the data will suffer a catastrophic loss. Organizations might be willing to accept the increased storage costs associated with higher storage tiers as value rises because the higher storage levels are likely to be better protected.[7] As value drops, the data can be migrated to a lower tier or perhaps even deleted entirely. Hence, the key to managing the cost of big data is to physically move or transfer it to a higher or lower tier each time there is a significant jump or decline, respectively, in its value.

Given the difficulties of discovering value changes, organizations have resorted to value proxies such as date of last use or frequency of use. Data governance practices can use these simple metrics to indicate an up or down shift in data value. Ambiguities can still arise when managers are asked to specify standards to embed in data governance practices—for example, if certain data is now used weekly instead of monthly, is that sufficient grounds to move it from tier 2 to tier 1, or is it better to wait until the data is used daily? In addition, the same rules might not apply to all data—rules for migrating email, for example, can vary from other types of data such as payroll records or monthly sales commission reports.

Table 1. Data governance practices.

Structural practices
• User involvement in policy setting and evaluation
• Steering committees or other means to assess data value and costs
• Creation of data ownership/stewardship rights and responsibilities
Operational practices (span all stages of the information life cycle)
• Enact retention policies/possible autodeletion or archiving of data
• Determine backup and recovery practices and parameters
• Establish and monitor access rights
• Set service level for protection of different types of information
• Establish chargeback procedures to recover operating costs
• Monitor risk factors that contribute to value-at-risk
• Migrate data between storage tiers based on defined criteria
• Plan and provision storage capacity
• Establish e-discovery and archival procedures
Relational practices
• Educate users and non-IT managers regarding storage utilization and costs
• Develop communications regarding policy effectiveness and user needs

DATA GOVERNANCE PRACTICES

Data governance practices are organizational policies or procedures that describe how data should be managed throughout its useful economic life cycle. These practices comprise three separate categories: structural, operational, and relational.[8]

Structural practices identify key IT and non-IT decision makers and their respective roles and responsibilities regarding data ownership, value analysis, and cost management. Structural practices can, for example, state the role of corporate legal counsel in setting policies and standards for protecting and using data. They also can specify committees to oversee compliance with internal policies or with legal rules about data retention.

Operational practices are the means by which organizations execute data governance. As Table 1 shows, these practices span a series of actions involving data migration, data retention, e-discovery, access rights, backup and recovery procedures, storage tiers, and cost allocation through chargebacks.

Organizations can combine different operational practices based on their specific needs and circumstances. Relational practices describe the formalized links among the CIO, business managers, and data users in terms of knowledge sharing, value analysis, education, training, and strategic IT planning.

A study of data governance practices at 30 large US organizations including Intel, Visa, Google, Johns Hopkins Medicine, AT&T, the American Red Cross, Limited Brands, Alaska Air, AstraZeneca, and Nissan uncovered evidence of all three types of data governance practices, although the extent to which each type of practice had been adopted varied widely.[7] Given the nature of their business, these organizations have a growing big data problem, yet their success with using IT in other contexts to improve performance reveals there is no magic bullet solution when it comes to managing big data.

Although data governance practices are instrumental to understanding and mitigating risk factors that can put data at risk, such as storage media failures or improper backups, they also are instrumental in shaping intermediate performance measures that can directly

Table 2. Data governance enablers and inhibitors.

Enablers	Inhibitors
Organizational factors	
Highly focused business strategy	Complex mix of products and services
Aligned IT and business strategy	Strategic misalignment
Centralized IT and organization structure	Decentralized IT and organization structure
Industry factors	
Regulations	Regulations vary by region (US, EU, and so on)
Predictable rate of data growth	Absence of industry-wide data standards
Technological factors	
Culture of promoting strategic use of IT	Packrat culture (data hoarding)
IT standardization	Legacy IT systems (weak integration)

increase the bottom line. These outcomes are industry-specific but reflect results such as fewer medical errors, lower expenses, and shorter hospital stays in the healthcare sector; improved pricing decisions in the energy trading sector; enhanced customer satisfaction in the automotive sector; and real-time decision making concerning scheduling, market analysis, and ticket pricing in the airline industry.

Data governance enablers and inhibitors

Various technical and organizational factors can impact the struggle to design and implement the data governance practices that can help organizations manage the value, risk, and cost of big data. Some of these factors can facilitate and accelerate the adoption and diffusion of data governance but others can act as inhibitors, slowing its adoption and use or perhaps preventing it outright.

Organizations that have a focused business strategy tend to be more prepared to enact a standard set of data governance policies; those that are more diverse in their strategic orientation might struggle to find policies that are equally relevant to each aspect of their business. For example, an organization espousing a low-cost strategy might try to minimize its storage spending, whereas one with a more customer-oriented focus might prefer to spend more on managing data for certain customer groups.

Organizations that can align IT and business strategy are more likely to agree on data governance because this alignment is often achieved through close cooperation between CIOs and business executives. Centralization can also play an enabling role as there is less need to allow business functions to enact their own policies or to try to reconcile potentially conflicting rules. Industry regulations push organizations to consider an appropriate system of policies rather than risk noncompliance penalties. IT standardization or data integration can make it even easier to devise common governance rules.

Inhibitors of data governance come in various forms, but foremost among them are legacy systems that force organizations to manage multiple, disparate, and disconnected systems. It is more difficult to devise practices to protect and manage data when it is duplicated across the organization. Deduplication can help limit this, but it does not solve the problem. A decentralized organization can find it difficult to create a common set of data governance policies or standards that will satisfy all decentralized users equally. For example, limits on data retention for critical applications might work well for some users but not for others.

Certain user behaviors can also limit the adoption of data governance. A pervasive packrat culture that fails to discourage the needless retention of data can be difficult to reverse. Retention of unstructured data in the form of user-generated images, video or music files, word processing documents, spreadsheets, and presentations is a leading cause of the data glut in organizations, accounting for some 60 percent of the total volume of big data.[1]

The breadth and complexity of an organization's product and service offerings can also inhibit the adoption of data governance. A multidivisional organization structured along product or market lines can complicate the adoption of standard data governance policies, with each division seeking exemption from a standard set of organization-wide policies by claiming that standard policies are incompatible with their idiosyncratic business needs. Where organizations have bowed to internal pressure to allow divisions to opt out of established firm-wide policies in favor of personalized policies, the result has been a chaotic mix of complex and in many cases contradictory policies. Users learn how to game the system when they feel that policies are inconsistent.

Table 2 highlights the various enablers and inhibitors of data governance based on prior research findings. While the enablers of data governance can vary by industry, inhibitors apply equally across all industries.

Data governance limitations

Although it is seen as inherently beneficial in the search for increased value from big data,[9] research also shows that there are limits to how much value can be obtained from data governance.[10]

Beyond a certain point, stricter data governance can have counterproductive effects. If employees cannot complete tasks without encountering data governance limitations or if they perceive governance policies to be

unjust, unnecessary, or excessive, there is a risk that they will develop risky workarounds to complete their assignments.

Intel addresses this governance concern through a "protect to enable" policy. The intent is not to scale back governance by removing onerous policies (real or imagined) but rather to recognize that innovation can sometimes depend on smart people doing unimagined things with data that could yield superior results. It is difficult to write data governance policies that encompass all eventualities or possible uses for data. Instead, Intel accepts the need to protect its information assets in a way that will enable those assets to be used to enhance innovation.

Data can only create value when it is used, so the intent is to motivate use within a framework of safeguards that seek to protect that value at all times. Although risk is still a valid consideration, there is an equal realization that it is impossible to protect against all risks. Consequently, organizations might need to be comfortable with being uncomfortable. The temptation to protect against all forms of technical and organizational risk could limit data use to the extent that it becomes difficult to derive increased value from data.

Setting data governance policies that help strike a balance between what are acceptable and affordable risks and how those policies can unlock value from data is the true key to developing and sustaining an effective data governance regimen.

GOVERNING NEW FORMS OF BIG DATA

The numbers describing the current and future growth rates of big data are truly staggering. While much of this growth is predicted to be unstructured—sensor data, social networking data, high-resolution images as used in DNA analysis, and video, for example—there is a growing awareness that much of the value of its retention will come from unexpected and somewhat questionable combinations of data.

The analysis of real-time streaming data such as the movement of customers through a retail store is likely to create large quantities of data and offer the potential for some novel insights, but it also creates some socioethical concerns that have repercussions for how these new forms of data will be governed. Regulations can vary from country to country, often putting up barriers to how data can be used. For example, France and other members of the EU have long debated whether retaining an IP address constitutes a breach of privacy. Global organizations might face a complex network of often contradictory country-specific laws that describe what they can or cannot do.

Regardless of the limitations on data use in each location, data governance policies might need to be infused with legal protections for multiple parties—customers, suppliers, employees, and business partners. A policy designed to secure supplier data might be insufficient or unmanageable for customer data. Hence, corporate legal counsel is increasingly proactive in creating, approving, and monitoring data governance policies for all user groups. While the applications that generate the data tend to identify the data owners, legal counsel can set limits on how data can be used. New forms of data and high-speed analytics can test these limits.

The extent to which an organization can act upon insights that come from new forms of data is still an unresolved issue, as the retail giant Target discovered in 2012, when it mailed marketing materials to a teenager whose parents did not know of her pregnancy.[11] Data governance policies must be sensitive to idiosyncratic privacy concerns despite any general policy limitations.

Policies that strike a balance between acceptable and affordable risks are the key to an effective data governance regimen.

Effective management of big data could lead to tremendous opportunities for organizations even as the technical challenges of managing unstructured data increase. When organizational reputation and the ethical nature of critical business processes become a concern, organizations may need to enact further levels of data governance to contain new forms of risk that are concurrent with managing new and emerging forms of data.

Socioethical issues also arise with the analysis of social networking data. Google, for example, can probe detailed epidemiological factors pertaining to the spread of infectious diseases based on search engine terms. Growing acceptance of BYOD (bring your own device) practices dull the line between work and home use of technology, potentially allowing organizations to monitor and analyze offsite use of their technology.

Data governance practices have a social and, in some cases, legal responsibility to safeguard personal data that, if compromised, could threaten personal privacy. Although it might be anathema to some organizations, it might be appropriate to involve a cross section of customers and other interested groups outside the organization in setting safeguards to protect data.

It is not unusual to find situations in which data loss or a privacy breach provides the incentive to enhance or, in some cases, enact data governance. Reactive approaches to data governance are not recommended because data is increasingly strategic, and data-oriented black swan events could put the entire organization at risk.

For data governance to be truly effective, it must be planned with a view to minimizing events that put data, and indeed the survival of the organization, at risk.

Data governance is no longer a luxury for organizations nor is it something they can implement with minimal effort. As IT hardware and software recede into the background as commodities and as information emerges as a strategic business asset—albeit with zero valuation for accounting purposes—it is imperative for organizations to spend time thinking about data governance and the range and scope of practices available to them.

Because the value of data changes over time and storage innovation is likely to change the economics of how data can be managed, the scale and scope of data governance practices might need to be changed over time as well. Despite the fact that some commentators ridicule the big data label while others brand it as overblown, the fact remains that data or information management is increasingly strategic.

Organizations that have the ability to unlock value from their data faster than their competitors will likely be the winners in the race to see who can get the most benefit from big data. History will likely reveal that data governance practices proved instrumental in this race.

References

1. C. Beath et al., "Finding Value in the Information Explosion," *Sloan Management Rev.*, vol. 53, no. 4, 2012, pp. 18-20.
2. P.P. Tallon and R. Scannell, "Information Life Cycle Management," *Comm. ACM*, vol. 50, no. 11, 2007, pp. 65-69.
3. "Downtime, Outages, and Failures: Understanding Their True Costs," blog, 6 Jan. 2013; www.evolven.com/blog/downtime-outages-and-failures-understanding-their-true-costs.html.
4. D.R. Merrill, *Four Principles for Reducing Total Cost of Ownership*, Hitachi Data Systems, 2011.
5. P. Weill, S.L. Woerner, and H.A. Rubin, "Managing the IT Portfolio (Update Circa 2008): It's All about What's New," *MIT Center for Information Systems Research* (CISR), vol. 8, no. 2B, 2008, pp. 1-4.
6. T.H. Davenport, P. Barth, and R. Bean, "How 'Big Data' Is Different," *Sloan Management Rev.*, vol. 54, no. 1, 2012, pp. 43-46.
7. P.P. Tallon, R.V. Ramirez, and J.E. Short, "The Information Artifact in IT Governance: Towards a Theory of Information Governance," to appear in *J. MIS*, Loyola Univ. Maryland, 2013.
8. K. Weber, B. Otto, and H. Österle, "One Size Does Not Fit All: A Contingency Approach to Data Governance," *J. Data and Information Quality*, vol. 1, no. 1, 2009, pp. 1-27.
9. V. Khatri and C.V. Brown, "Designing Data Governance," *Comm. ACM*, vol. 53, no. 1, 2010, pp. 148-152.
10. M. Harkins, *Managing Risk and Information Security: Protect to Enable*, Apress Media, 2013.
11. C. Duhigg, "How Companies Learn Your Secrets," *The New York Times*, 16 Feb. 2012.

Paul P. Tallon *is an associate professor and director of the David D. Lattanze Center for Information Value at the Sellinger School of Business, Loyola University Maryland. His research interests include the economic and organizational impacts of IT, strategic alignment, IT portfolio analysis, and the economics of data management. Tallon received a PhD in management from the University of California, Irvine. Contact him at pptallon@loyola.edu.*

Selected CS articles and columns are available for free at http://ComputingNow.computer.org.