IBM Center for
The Business of Government

Using Technology Series

Realizing the Promise of Big Data

Implementing Big Data Projects

Kevin C. Desouza
Arizona State University

Using Technology Series 2014

Realizing the Promise of Big Data: Implementing Big Data Projects

Kevin C. Desouza
Associate Dean for Research, College of Public Programs
Associate Professor, School of Public Affairs and
Interim Director, Decision Theater
Arizona State University

IBM Center for
The Business of Government

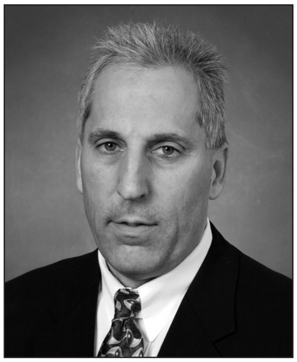
Table of Contents

Foreword

On behalf of the IBM Center for The Business of Government, we are pleased to present this report, *Realizing the Promise of Big Data: Implementing Big Data Projects,* by Kevin Desouza, Arizona State University.

Daniel J. Chenok

Professor Desouza provides a clear and useful introduction to the concept of big data, which is receiving increasing attention as a term but also lacks a commonly understood definition. In describing big data, Desouza writes, “Big data is an evolving concept that refers to the growth of data and how it is used to optimize business processes, create customer value, and mitigate risks.” Desouza also describes the differences in the use of big data in the public and private sectors.



Gregory J. Greben

A key contribution of the Desouza report is his descriptions of how big data is being used in federal, state, and local government today. His examples include the Internal Revenue Service, the state of Massachusetts, and the New York City Business Integrity Commission.

Over the last year, Professor Desouza conducted extensive interviews with chief information officers (CIOs) across the United States at the federal, state, and local level. The goal of the interviews was to better understand the implementation challenges facing CIOs and their organizations as they undertake big data projects. Desouza presents 10 key findings from his interviews. He also presents detailed descriptions of the three key stages in implementing a big data project: planning, execution, and post-implementation.

This report continues the IBM Center's interest in the concepts of big data and analytics. In cooperation with the Partnership for Public Service, the IBM Center for The Business of Government recently issued its third report on analytics: *From Data to Decisions III: Lessons from Early Analytics Program.* The use of analytics in agency program management and decision-making relies on big data.

Collectively, this report and the Partnership-IBM Center's three reports on analytics provide a roadmap for federal leaders on how to move to the new era of big data and analytics. We hope these reports will assist all public managers as they continue their quest to better use data that they collect on an ongoing basis.

Daniel J. Chenok
Executive Director
IBM Center for The Business of Government
chenokd@us.ibm.com

Gregory J. Greben
Vice President
Business Analytics & Optimization
Practice Leader, IBM U.S. Public Sector
greg.greben@us.ibm.com

Executive Summary

Today, public agencies operate in data-intensive environments. Individuals and organizations generate and disseminate data at an ever-increasing rate. Due to advances in computing and communications technology, the costs associated with creating real-time data on a wide variety of objects, agents, and events have fallen drastically. Mobile technologies, online social networks, and social media are ingrained into our community and professional fabrics. Not surprisingly, public agencies have had to direct their attention to the opportunities and challenges of operating in a big data environment.

Big data is a concept that has exploded in recent years, due largely to our ability to store and analyze massive amounts of data efficiently and cost-effectively. Data is created every second of every minute of every hour; we now create 2.5 quintillion bytes of data per year. That is 2,500,000,000,000,000,000 bits of information. Private-sector organizations including Amazon, Target, and Facebook have made millions of dollars using big data analytics. The International Data Corporation expects the business intelligence and analytics market to grow to nearly $50 billion by 2016.[1] The public sector has taken notice of big data's power and has sought its value in solving social challenges and creating better citizen service delivery.

In 2012, the Obama administration announced the Big Data Research and Development Initiative, which provided more than $200 million to launch new big data projects across six federal departments and agencies with the goal of accelerating scientific discovery, improving national security, enhancing teaching and learning, and increasing innovation. States, cities, and counties, both within the United States and beyond, are also experimenting with a wide assortment of big data programs.

Big data is a new frontier for the public sector. It has captured the attention of public managers across the globe. Agencies realize that their datasets represent critical resources that need to be managed and leveraged. Public sector use of big data and big data analytics is wide-ranging; some organizations have no experience with big data, while others have taken on small to moderate-sized projects. Drawing on interviews with chief information officers (CIOs) from every level of government (federal, state, and local), this report presents implementation steps grouped by the phases of a big data project:

- Planning
- Execution
- Implementation

Each step holds individual value, but also contributes to an overall framework for managing public-sector big data projects. The key steps serve as best practices for CIOs and public man-

[1] Dan Vesset, Brian McDonough, Mary Wardley and David Schubmehl. Worldwide Business Analytics Software 2012–2016 Forecast and 2011 Vendor Shares. Idc.com. June 2012.

agers to consider for their agency as they embark on big data projects. Some agencies are struggling to realize the promise of big data. Big data is a wide-ranging and somewhat elusive term that has left many CIOs grasping for a better understanding.

CIOs now seek the most cost-effective, innovative ways to deploy big data given limited funding and diminishing intellectual capital. CIOs are struggling to recruit and attract data scientists equipped to handle big data analytics. Additionally, the pervasive issues of privacy and ethical data use remain at the forefront of leveraging big data. Yet, even given these challenges, CIOs are innovating and finding ways to launch big data programs. They realize their data reservoir's value as a critical asset that should be leveraged to enable evidence-based decision-making, streamlining of operations, and increased value of citizen services.

In the next few years, nearly all public agencies will grapple with how to integrate their disparate data sources, build analytical capacities, and move toward a data-driven decision-making environment. Big data is increasing in importance for public agencies, and big data programs are expected to become more prominent in the near future. Through the use of big data, analytics now holds great promise for increasing the efficiency of operations, mitigating risks, and increasing citizen engagement and public value.

The report contains the following 10 findings:

- **Finding One:** Public agencies are in the early days of their big data efforts.
- **Finding Two:** Many CIOs fight the perception that big data is a passing fad.
- **Finding Three:** Most CIOs are now primarily dealing with the issue of managing large volumes of data, integrating data across database systems, and building an analytical capacity to mine data.
- **Finding Four:** CIOs report that some big data projects are now focused on streamlining business processes.
- **Finding Five:** CIOs do not anticipate significant investments in technology.
- **Finding Six:** CIOs report a need to bolster their human capital, including their analytical capability.
- **Finding Seven:** CIOs are now exploring approaches to data governance.
- **Finding Eight:** CIOs do not recommend IT units as owners of big data projects.
- **Finding Nine:** CIOs believe that collaborative leadership is crucial for the success of big data projects and recommend the creation of working groups to oversee projects.
- **Finding Ten:** CIOs are becoming champions of analytics and evidence-driven decision-making.

The box on the following page, *Key Steps in Implementing a Big Data Project,* presents the report's key steps, detailed in the concluding section of the report.

Key Steps in Implementing a Big Data Project

Stage One: Planning

Step One: Before undertaking a big data project, chief information officers (CIOs) must do their homework, which includes understanding relevant business and legal policies.

Step Two: Before undertaking a big data project, CIOs must build a coalition that includes reaching out to peers.

Step Three: In communicating big data projects, CIOs must define the broader opportunity.

Step Four: In selecting a big data project, CIOs should begin with the lowest-hanging fruit.

Step Five: In starting a big data project, CIOs must ensure strategic alignment of the project. This includes lining up sponsors.

Step Six: In leading a big data project, CIOs should be privacy advocates.

Step Seven: Use taskforces in implementing a big data project.

Step Eight: CIOs should outline expected resistance and plan for it when undertaking a big data project.

Step Nine: CIOs should develop key performance indicators before starting a big data project.

Step Ten: Design a risk mitigation plan before starting a big data project.

Stage Two: Execution

Step Eleven: When a big data initiative is underway, CIOs should constantly gauge the pulse of the program.

Step Twelve: During the big data initiative, it is crucial to communicate, communicate, and communicate.

Step Thirteen: Throughout a big data project, CIOs must manage scope creep.

Step Fourteen: During a big data project, CIOs must stay focused on the data and not get caught up or distracted by the technologies involved in the project.

Step Fifteen: If necessary, CIOs should pull the plug on a big data project.

Stage Three: Post-Implementation

Step Sixteen: At the end of a big data project, conduct a postmortem and an impact analysis.

Step Seventeen: Identify the next project.

Understanding Big Data

> *"Big data is indeed a big deal."*
>
> — Dr. John Holdren
> Director, Office of Science and Technology Policy,
> White House

The Opportunity Ahead

Technological advancements have made it easier to collect and store data. We are generating and storing data on a nearly pervasive basis and across multiple environments including work and home. The cost of storing data has also fallen sharply. Storage devices are not only cheaper, but there have been significant advancements in the science of databases and information retrieval.

We are living in a data-rich world, and organizations must be apt at turning data into insights for evidence-based decisions.[2] More and more organizations understand that the insights provided by data are only useful if they can be properly communicated and lead to data-based decision-making.[3]

Big data success stories are plentiful in the private sector. In July 2012, Merck, maker of the allergy pill Claritin, used specialized weather forecasts that offered historical data, as well as current weather reports, to anticipate hay fever in May 2013. Merck analysts knew that many allergens would remain dormant throughout March and April due to unseasonably cold weather. They concluded that May's warmth would result in higher than average pollen counts and increased demand for allergy medication. Merck used this information when developing business strategies and promotional material; a partnership with Wal-Mart Stores, Inc. was established and personalized promotions were created based on zip code data—resulting in increased revenue for the quarter. [4]

General Motors invested in telematics for its OnStar technology. Telematics is the blending of computers and wireless telecommunications technologies, ostensibly with the goal of efficiently conveying information over vast networks to improve a host of business functions or government-related public services. The term has evolved to refer to automobile systems that combine global positioning satellite (GPS) tracking and other wireless communications for automatic roadside assistance and remote diagnostics.

[2] Thomas Davenport. Data is Worthless if You Don't Communicate It. Blogs.hbr.org. June 18, 2013. http://blogs.hbr.org/2013/06/data-is-worthless-if-you-dont/.

[3] Jeff Bertolucci. Become A Data Scientist ... In 12 Weeks? Informationweek.com. September 18, 2013. http://www.informationweek.com/big-data/news/big-data-analytics/become-a-data-scientist—in-12-weeks/240161283.

[4] Kim S. Nash. How to Profit From the Ultimate Big Data Source: The Weather. CIO.com. May 24, 2013. http://www.cio.de/index.cfm?pid=156&pk=2916885&p=1.

OnStar uses telematics through GPS tracking and wireless communication to provide driver assistance, including vehicle security, information and diagnostic services, and data capture in case of an accident. Using telemetry data drawn from users, OnStar and GMAC Insurance Group joined forces to create a program that offers customers who drive fewer miles lower insurance premiums—resulting in increased customer satisfaction.[5] In the private sector, top-performing companies use analytics five times more than poorly performing companies do, and make decisions based on rigorous analysis at a rate more than double that of lower performers.[6] Big data and analytics represent an emerging opportunity for public institutions to streamline business processes, increase citizen engagement, innovate, and embrace evidence-driven decision-making. For public managers, big data represents an opportunity to infuse information and technology into the design and management of organizations, personnel, and resources.

Efforts in fields such as data mining and information visualization have given us the resources necessary to parse and traverse large collections of data that range from online, basic, and free to more robust and expensive.[7] Leading organizations are investing in their human capital to prepare themselves to take advantage of operating in data-intensive environments.[8] According to the McKinsey Global Institute report on big data, over 1.5 million trained big data managers are needed to take advantage of all the data we generate.[9] The term “data scientist” has entered our vernacular and many have touted this as a critical skill set that organizations will seek.[10] Some, including Tom Davenport, have called it the “sexiest job of the 21st century.”[11]

Big data and analytics represent an emerging opportunity for our public institutions to streamline business processes, increase citizen engagement, innovate, and embrace evidence-driven decision-making. For public managers, big data represents an opportunity to infuse information and technology into the design and management of organizations, personnel, and resources.

What is Big Data?

The term big data emerged in the 2000s as a rhetorical nod to vast, constantly increasing amounts of data.[12] Big data is an evolving concept that refers to the growth of data and how it is used to optimize business processes, create customer value, and mitigate risks. In the book *Big Data: A Revolution that Will Transform How We Live, Work, and Think,* authors Viktor Mayer-Schonberger and Kenneth Cukier write that one way to think about big data is that “big data refers to things one can do at a large scale that cannot be done at a smaller one, to extract new insights or create new forms of values, in ways that change markets, organizations, and the relationships between citizens and governments, and more.”[13]

[5] Justin Bachman. OnStar: GM’s Not-So-Secret Weapon. BusinessWeek.com. May 30, 2013. http://www.businessweek.com/articles/2013-05-30/onstar-gms-surprisingly-strong-toehold-in-the-tech-world

[6] Michael S. Hopkins, Steve LaValle, Eric Lesser, Rebecca Shockley, and Nina Kruschwitz. Big Data, Analytics and the Path from Insights to Value. 2011. *MIT Sloan Management Review,* Issue 52, Volume 2.

[7] Genie Stowers. *The Use of Data Visualization in Government.* IBM Center for The Business of Government. Businessofgovernment.org. 2013.http://www.businessofgovernment.org/sites/default/files/The%20Use%20of%20Visualization%20in%20Government.pdf.

[8] Thomas H. Davenport and D.J. Patil. Data Scientist: The Sexiest Job of the 21st Century. Hbr.org. October, 2012. http://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/ar/1.

[9] James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, Angela Hung and Byers. Big Data: The Next Frontier for Innovation, Competition, and Productivity. McKinsey.com. May 2011. http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation.

[10] Thomas H. Davenport and D.J. Patil.

[11] Thomas H. Davenport and D.J. Patil.

[12] Steve Lohr. How Big Data Became So Big. NYtimes.com. August 11, 2012. http://www.nytimes.com/2012/08/12/business/how-big-data-became-so-big-unboxed.html?_r=0

[13] Viktor Mayer-Schonberger and Kenneth Cukier. *Big Data: A Revolution that Will Work Transform How We Live, Work, and Think.* Houghton Mifflin Harcourt, 2013.

Much of big data's growth has come from necessity. Data has been stored in various forms throughout history. For example, in 1880, the United States conducted a national census of 50 million people that collected demographic information including age, gender, number of individuals in the household, ethnicity, birth date, marital status, occupation, health status (blind, deaf, insane, disabled), literacy, and place of origin. All of this was logged by hand, microfilmed, and sent to be archived in state archives, libraries, or universities. Seven to eight years were required to properly tabulate Census data after initial collection efforts.

In 1890, the Census Bureau streamlined its data collection methods by implementing machine-readable punch cards that could be tabulated in one year.[14] In the most recent U.S. Census conducted in 2010, the Bureau used a range of emerging technologies to survey the populace.[15] The Census Bureau currently employs geographic information systems, social media, videos, intelligent character recognition systems, and sophisticated data-processing software. These technologies play a key role in establishing an effective communication mechanism that reaches nearly every citizen in the nation. Data is being generated across a multitude of technologies on an almost pervasive basis.

While it is hard to trace the origins of big data in the public sector to a definite event, one could argue that it came to the forefront in the aftermath of the terrorist attacks on September 11, 2001. In *The 9/11 Commission Report,*[16] the Commission identified the U.S.'s inability to "connect the dots" between various pieces of available data to prevent the terrorist attack.[17] This led to the creation of the Office of the Director of National Intelligence that was to oversee the collection, analysis, interpretation, and implementation of actions on strategic targets (e.g., Iraq, Iran, and North Korea). News about big data programs from the U.S. intelligence community continues to this day, with the most recent being a series of news stories about programs at the National Security Agency (NSA).

A series of Vs describes the dimensions of big data:

- *Volume* considers the amount of data generated and collected
- *Velocity* refers to the speed at which data is analyzed
- *Variety* indicates the different types of data that are collected
- *Viscosity* measures resistance to flow of data[18]
- *Variability* measures the change rate of flow
- *Veracity* measures biases, noise, abnormality, etc.
- *Volatility* indicates how long data is valid for and should be stored for

While all Vs are increasing, they are not equal. For example, consider the case of *volume*: given that the processing capability of hardware devices continues to grow exponentially, data volume is not likely to cause major issues. Data that could not be stored on large, fixed hard drives several years ago can fit on mobile devices today. The world's data is doubling every 18

[14] Such machine readable cards became the foundation on which companies like IBM were developed.

[15] Kevin Desouza and Akshay Bhagwatwar. Leveraging Technologies in Public Agencies: The Case of the U.S. Census Bureau and the 2010 Census. *Public Administration Review,* 72 (4), 2012, 605–614.

[16] *The 9/11 Commission Report.* http://www.gpo.gov/fdsys/pkg/GPO-911REPORT/pdf/GPO-911REPORT.pdf

[17] Steve Vinsik. 9/11, A Decade Later—Connecting the "Big Data" Dots: A Decade of Lessons Learned. GSNmagazine.com. September 16, 2011. http://www.gsnmagazine.com/article/24545/911_decade_later_connecting_%E2%80%98big_data%E2%80%99_dots_decade.

[18] Ray Wang. Monday's Musings: Beyond The Three V's of Big Data—Viscosity and Virality. Softwareinsider.org. February 27, 2012. http://blog.softwareinsider.org/2012/02/27/mondays-musings-beyond-the-three-vs-of-big-data-viscosity-and-virality/.

months,[19] presenting the public and private sectors with new opportunities to transform information into insight. As the volume of data increases along with the tendency to store multiple instances of the same data across varied devices, the science of information search and retrieval will have to advance.

The most challenging *V* for organizations is *variety*. Organizations have built information systems to tackle data elements in specific categories. As an example, an HR system processes data on employees, while a CRM holds data on customers. The challenge facing many organizations is to find economical ways of integrating heterogeneous datasets while allowing for newer sources of data (both in terms of origin and type) to be integrated within existing systems; for example, integrating data from social media platforms with data extracts from traditional CRM systems. Ensuring the data collected is of sufficient *veracity* is also critical. Today, due to the proliferation of social networks and social media, much of the data being collected needs to be thoroughly analyzed before decision-making, as the data can be easily manipulated. Consider platforms such as Twitter, where fake Twitter accounts can generate content that might seem legitimate, and even mobilize action in the online community.

In an article published in *Wired* magazine, Chris Anderson argues that the difference between the Petabyte Age and other byte movements is organizational capacity. He notes that "kilobytes were stored on floppy disks. Megabytes were stored on hard disks. Terabytes were stored in disk arrays. Petabytes are stored in the cloud. As we moved along that progression, we went from the folder analogy to the file cabinet analogy to the library analogy to—well, at petabytes we ran out of organizational analogies."[20] In the context of big data, chances are low that any one organization will own all the data it needs or can solely rely on only data it has. Hence, we must think in terms of data that resides across networks. These networks can span organizations, systems, and individuals and bridge the public, private, and nonprofit spaces. In addition to data residing across networks, the expertise to analyze and visualize the data rarely resides in one organization. The need to develop viable collaborative partnerships for leveraging big data is becoming critical.

The newfound ability to collect data about a target population at an affordable cost has allowed researchers to avoid sampling bias. Some have argued for the irrelevance of theory testing today, as large quantities of data about the population can be mined directly. The data can speak for itself, or, more specifically, analyses can speak for themselves.[21]

A point of caution is worth noting here. While big data is available in abundance in the physical sciences (e.g., weather, astronomy, energy, etc.), the social sciences still suffer from limited data in most cases. This is especially true of some of the world's most complex social challenges. Consider the case of human trafficking. Clearly, this is a complex problem on the radar of a wide array of stakeholders. Countries, NGOs, the private sector, and even individuals are contributing resources to combat this global challenge. Yet the underlying data being used to drive policy making on this challenge is sparse, incomplete, and messy. The challenge for the social scientist is to learn from peers in the physical sciences about assembling, curating, and analyzing datasets to create truly big datasets. Most public agencies do not have datasets that meet the traditional definition of big data when we think of the "volume" dimension. However, public agencies do have the opportunity to curate datasets that can make a difference in advancing their mission. This can be achieved through building collaborative alliances around

[19] Matt Prigge. Data Storage: Buried Alive! Infoworld.com. December 7, 2009. http://www.infoworld.com/d/data-explosion/data-storage-buried-alive-037.

[20] Chris Anderson. The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. Wired.com. June 23, 2008. http://www.wired.com/science/discoveries/magazine/16-07/pb_theory/.

[21] Nizar Diamond Ali. *Petabyte Age.* Ezinearticles.com. January 28, 2010. http://EzineArticles.com/3644450.

data sharing across a wide array of sectorial networks, while tapping into expertise to analyze and mine the data.

Differences in Big Data in the Public and Private Sectors

There are many differences in how the public and private sectors are implementing big data projects. The goals and rules guiding each sector are different, making it difficult if not impossible to easily transfer tried-and-true private-sector practices to the public sector. Four major differences in how the public and private sectors handle big data are presented next.

Access, control, and data. In the private sector, most organizations have become skilled in their ability to collect immense amounts of information on individuals with limited pushback. For example, when using a mobile phone or a particular mobile app, customers agree—through acceptance of the app's terms and conditions—to allow the developer to collect information on them in real time, especially when facilitated through technology. Private-sector data collection is so ubiquitous that customers have come to expect it. Many apps, such as mobile flashlights or mobile dictionaries, commonly store personal information that can

Who is Generating Big Data?

In the area of health care, big data certainly has room to improve process efficiency and patient outcomes. The health care industry is working to figure out the most appropriate and ethical ways to collect, manage, store, and analyze data before deploying big data initiatives. Big data initiatives could impact most areas of the health care industry, including billing, diagnostic tests, medication administration, quality and outcome indicators, and patient records management. In the health care space, we have regulations such as HIPAA that govern how information is collected, stored, shared, and analyzed. Today, a massive amount of data is being generated on an individual's health and is not governed by regulation.

Individuals are creating and disseminating personal health information in real time using myriad mobile apps and technologies. Take, for instance, the NikePlus Fuel Band SE. It is advertised for its ability to track "your active life" which includes running and walking habits. NikePlus has entered into partnerships with other apps such as Sprout, which allows users to integrate their Fuel Band data into an employer-sponsored fitness program, or HighFive, which offers discounts and coupons on achieving fitness goals. While the benefits are clear, there are notable negative aspects to such tracking. Most notably, consumers might not appreciate giving data about daily habits, their whereabouts, or their personal level of fitness because it could potentially be used against them. While individuals are provided with an easy way to manage their fitness, track activity, and even record workouts, the same apps provide the creators with valuable information on the individual. Beyond the specifics of their health-related behaviors, apps tap into GPS data, contacts, activity on the web, and other information about the user.

Beyond actively sharing their health-related data through the use of mobile apps, individuals are sharing information on their health status on social networks, including updates on various social media sites to alert friends when they are feeling ill. In some cases, individuals get more specific and even share their reaction to medications used. All this data is ripe for being mined and analyzed to learn about the individual. In addition, given the fact that such information is time-stamped and in many cases geo-coded with location information, the targeting of advertisements can be quite precise.

To date, most legislation (e.g., HIPAA) dealing with the protection of sensitive medical and health-related data from patients has not kept up with the era of big data, in particular with regard to user-generated medical information in the public sphere.

include user location, gender, and a unique identifier of the mobile device. Others can store contact list information and pictures from photo libraries.

The federal government is aware of this issue and currently working on it. In February 2012, the White House released "Consumer Data Privacy in a Networked World: A Framework for Protecting Privacy and Promoting Innovation in the Global Digital Economy" (The Privacy Blueprint). The Privacy Blueprint established a Consumer Privacy Bill of Rights and directed the National Telecommunications and Information Administration in the Department of Commerce to convene multi-stakeholder processes to develop legally enforceable codes of conduct that specify how the Consumer Privacy Bill of Rights applies to privacy on mobile devices.

Consumer data purchasing. Private enterprises can easily purchase data from data banks, whether credit card bureaus or other data aggregation organizations. For instance, Amazon.com has over 152 million customers registered on their site. Individual buying patterns are tracked and targeted advertisements are generated based on user behavior. Instead of allowing other companies to purchase customer data, Amazon aggregates data and gives retailers the opportunity to bid on targeted ad space. Facebook allows advertisers to target potential customers by matching customers with their interests. For instance, Facebook allows for real-time, cookie-based retargeting of people who visit an advertiser's website, e.g., if a consumer visits auto dealers' web pages, when they log on to Facebook they will see ads for cars.

In the public sector, accessing large databases outside those in one's agency is challenging. Government entities face increased difficulty obtaining data if the needed information is only available through a partnership with a private-sector firm. The NSA received heavy criticism when it was reported that during a single day last year, the NSA's Special Source Operations division collected 444,743 e-mail address books from Yahoo, 105,068 from Microsoft's Hotmail, 82,857 from Facebook, 33,697 from Google's Gmail service, and 22,881 from unspecified other providers.[22] The controversy stirs long-standing sentiments among Americans, who view such programs as a threat to civil liberties.

Data collection efforts are largely influenced by shared perceptions of an organization's motivations. Websites and apps developed by private organizations are generally viewed as non-invasive because the data is used to increase the firm's competitiveness—an acceptable motivation. Some may appreciate such data collection because it often results in intuitive applications that are easier to use. Public sector data collection is viewed as invasive because it may appear to provide little benefit and can be used against citizens. In addition, evidence to date does not exist to show that the data collected results in improvements to governance, changes in civic society, or better living conditions. This attitude presents an opportunity for public agencies to do things differently and begin to engage in the big data space, while being mindful of citizens' concern for privacy and ethical data collection and usage.

Investments in information technology. The private sector has invested more heavily in information technology than the public sector has. In the private sector, information technology is seen as a critical element of being competitive and relevant in the marketplace. Across a wide spectrum of industries, there has been a growth in information technology expenditures. Information technologies have been infused across a myriad of business processes, and have even been used to reach new customers.

[22] Alistair Barr. Latest NSA Revelation is Black Eye for Yahoo. Usatoday.com. October 15, 2013. http://www.usatoday.com/story/tech/2013/10/15/nsa-yahoo-black-eye/2988599/.

To date, the public sector has not invested necessary resources in modernizing its information technology infrastructures and capabilities (i.e., hiring skilled workers and training). In the landmark launch of the Affordable Care Act, visitors to Healthcare.gov experienced trouble logging onto the site and creating accounts, slow page loads, and unexpected error messages. According to *USA Today,* the computer glitches can be attributed to a platform that was "built using 10-year-old technology that may require constant fixes and updates for the next six months and the eventual overhaul of the entire system."[23] To help move the issue forward, the Department of Health and Human Services hired additional IT professionals to help manage the system. Issues of data security on the platform have also been highlighted in light of the recent problems, and may undermine the use of the system when it is back up and running.[24]

Arcane information systems, especially those found in the public sector, have not been designed to handle the Vs. In 2000, the Federal Bureau of Investigation (FBI) launched the "Virtual Case File" (VCF) initiative, which was designed to offer a searchable database that would allow agents to "connect the dots" and follow up on disparate fragments of intelligence. An initial investment of $170 million was made in 2005; the project was an abject failure and was officially abandoned several years later. The failure of the VCF initiative led the Inspector General of the U.S. Department of Justice to conduct an official audit of the program. Twelve major issues contributed to the failure of the program:

- Unrealistic or unarticulated project goals
- Inaccurate estimates of needed resources
- Badly defined system requirements
- Poor reporting of the project's status
- Unmanaged risks
- Poor communication among customers, developers, and users
- Use of immature technology
- Inability to handle the project's complexity
- Sloppy development practices
- Poor project management
- Stakeholder politics
- Commercial pressures[25]

Many public organizations experience shortcomings, such as the FBI's in the VCF case, when initiating big data projects. It is crucial that public sector agencies develop IT project management abilities to ensure survival in the big data era.

Ethical and privacy concerns. Public agencies have to weigh the ethical and privacy concerns of big data more heavily than the private sector does. For all the good that big data does, there are negative aspects of its capabilities that public agencies cannot and should not ignore. As the amount of data grows, so does the ability to know more about citizens and the potential to use such information in a harmful or discriminatory way. Data discrimination has

[23] David Jackson. Obama to Criticize Health Care Website Problems. Usatoday.com. October 21, 2013. http://www.usatoday.com/story/news/2013/10/21/obama-health-care-internet-glitches-health-and-human-services/3142759/.

[24] Jose Paligry. Security Hole Found in Obamacare Website. CNN.com. October 29, 2013. http://money.cnn.com/2013/10/29/technology/obamacare-security/index.html?hpt=hp_t3

[25] Robert N. Charette. Why Software Fails. Spectrum.ieee.org. September 2, 2005. http://spectrum.ieee.org/computing/software/why-software-fails.

heightened with the explosion of big data analytics. Information from datasets that are wide-reaching and interpretable for a certain population of people can be harmful and lead to data discrimination.

Big data allows for the cross-referencing of information and can create detailed portraits of groups of people that could affect such issues as loan approval, life insurance qualification, or hiring. Much of this information can be purchased and used for a variety of reasons. Often, those being judged or analyzed by big data never have an idea what is happening with their information.

While most public managers see big data as a way to make the lives of their constituents better, there remains a large element of potential harm in its use. Public agencies in New York came under critical scrutiny for public disclosure for an uncalculated, and some might argue an emotional, response to a real-time situation. In the aftermath of the Sandy Hook Elementary School shooting in Connecticut, a group of researchers obtained through the Freedom of Information Act information regarding gun owners living in the suburbs of Westchester, Rockland, and Putnam counties in New York. The researchers published an article and interactive visual map with the gun owners' names and addresses. Although the information was published to "provide open knowledge" about local gun possession, it also provided valuable information to assist criminals, who could now target homes without gun owners or target homes to steal guns and sell them illegally. Traditional governing laws and unforeseen consequences double the challenges on public agencies when considering the appropriate ways to share, store, manage, and analyze data.

Another ethical concern associated with the growth in big data analytics is about privacy. The benefits of having data available for mining are clear, but there remain issues of privacy that many are rightly concerned with maintaining. As Alistair Croll notes, "Data doesn't invade people's lives; lack of control over how it's used does." The availability of a great deal of data is not new, but the ability to analyze it is, and it requires careful attention to stewardship. Most organizations have some concerns about privacy, albeit at differing levels. Traditionally, most organizations, public and private, undertake take various de-identification measures (anonymization, pseudonymization, encryption, key-coding, data sharding) to protect their constituents and stakeholders. Even the use of de-identification measures may not be enough. Computer scientists have shown on a number of occasions that anonymized data can often be re-identified and attributed to specific individuals, thus undermining good-faith efforts to protect individuals' privacy.[26]

[26] Omer Tene and Jules Polonetsky. Privacy in the Age of Big Data: A Time for Big Decisions. Stanfordlawreview.com. February 2, 2012. http://www.stanfordlawreview.org/online/privacy-paradox/big-data#footnote_7.

The Uses of Big Data

In the Private Sector

The H1N1 flu epidemic. The private sector is now tackling problems traditionally within the purview of public agencies. In 2009, H1N1 was a new human-to-human transmitted influenza virus for which no immunity existed. Seasonal influenza is a major health care concern, and new and aggressive strains of the flu (such as H1N1) can be particularly challenging for public health agencies, i.e., the Centers for Disease Control and Prevention (CDC). H1N1 was a particularly insidious strain of influenza; it spread rapidly and had the potential to cause numerous fatalities. Google supported early detection of H1N1 through the launch of Google Flu Trends, a tool that identified flu outbreaks by tracking flu-related search queries. Google Flu Trends employed large-scale data analytical techniques to predict the spread of the H1N1 virus through the use of near-real-time data tracking of health-seeking behaviors. Data collection involved online search queries and other indirect signals of influenza activity such as call volume to triage advice lines and over-the-counter drug sales. Google also used publicly available historical data from the CDC's U.S. Influenza Sentinel Provider Surveillance Network, which monitors influenza spread by employing national networks of physicians that report cases of influenza-like illness (ILI)—a diffuse set of symptoms, including high fever, which for purpose of analysis are used as a proxy for flu.

Google's early detection model processed hundreds of billions of individual searches from five years (2003–2008) of Google search logs, producing weekly counts for the most common search queries and separating those counts by state. In the CDC's case, physicians in their surveillance network were requested to report cases but information about the H1N1 outbreak was delayed due to several problems with reporting. First, the lag effect of symptoms caused physicians to misdiagnose or attribute the flu to other illnesses such as asthma, diabetes, chronic heart problems, and immune system problems.[27] Second, delays were present in the routing of information from the point of detection to central processors at the CDC. Finally, H1N1 detection relies heavily on laboratory confirmation and investigation—which is extremely resource-intensive. Millions of individuals with a mild form of H1N1 slowed detection and the actual number of severe cases remains unclear today.[28] Significantly, information extracted from both Google Flu Trends and the CDC's surveillance of ILI was a near-exact match. Google, however, was able to deliver their results days faster.

[27] Mike Stobbe. U.S. Swine Flu Cases May Have Hit 1 Million. Huffingtonpost.com. June 25, 2009. http://www.huffingtonpost.com/2009/06/25/us-swine-flu-cases-may-ha_n_221240.html.

[28] Changes in Reporting Requirements for Pandemic (H1N1) 2009 Virus Infection—Pandemic (H1N1) 2009 Briefing Note 3 (Revised). July 16, 2009. http://www.who.int/csr/disease/swineflu/notes/h1n1_surveillance_20090710/en/index.html.

In the Federal Government

The United States Postal Service. Public agencies are making strides in the big data space. Starting in 2006, the United States Postal Service (USPS) began using big data to prevent billions of dollars in losses, combat budget cuts, and increase efficiency. Processing more than 528 million pieces of mail each day—that's 6,100 mailings each second—the USPS has good reason to be an active participant in the big data revolution.

Unknown to most, USPS owns and operates one of the largest non-classified supercomputing databases on the planet. Located in Eagan, Minnesota, the Eagan IT and Accounting Service Center is the hub of supercomputing operations—data from each piece of scanned mail is processed at this facility. Data collected from scans is compared in real time to a database including approximately 400 billion records. Complex algorithms help carry out fraud detection and other validation tests on the data before it is routed to a delivery center. Average processing time for each item is 50 to 100 milliseconds. The overall goal of the system is to detect abnormalities that indicate fraud before suspicious mailings arrive at their intended local post office.

The United States Internal Revenue Service (IRS). An increasing number of companies and agencies are exploring new ways to use big data to help streamline processes and increase revenue. The IRS is no different. A bureau of the Department of the Treasury, the IRS is the revenue service of the United States government, and is responsible for collecting taxes and enforcing the Internal Revenue Code. The IRS collects over $2.4 trillion in taxes from nearly 250 million tax returns each year.[29] The IRS wants to use big data analytics of financial and social information to address taxpayer error, evasion, and other sources of lost revenue.

Taxation is a heavily discussed topic in politics, and leaders face great pressure to find innovative ways to reduce revenue losses. The IRS reportedly loses an estimated $300 billion each year in taxpayer error or cheating tactics. To combat such loss, the IRS will use big data analytics to combine already-collected taxpayer information with new social information on individuals' digital activities, such as credit card payments, e-pay transactions, eBay auctions, and Facebook posts. The IRS already has access to personal data on Americans, such as Social Security numbers, health records, and credit card transactions. It is now seeking more social information to conduct "robo-audits."

Robo-audits will allow the IRS system to track citizens' online activity and flag patterns of concern. Collection and analysis of such data will allow the IRS to generate and track unique attributes regarding financial behaviors. Much of the data will be used for research, but it will also aid in tax enforcement and combat noncompliance. In the past, such third-party (social) data has only been used if irregular returns merited more attention.

The Bureau of Conflict and Stabilization Operations in the Department of State. Big data is important to U.S. foreign policy. In January 2012, the Bureau of Conflict and Stabilization Operations (CSO) was established by the Department of State to provide effective and coherent crisis response overseas. Specializing in transitional societies, CSO works with countries in need of direct intervention; these countries can be at risk of violent extremists, weapons proliferation, government corruption, and organized crime. Supporting this type of conflict and crisis response is done through the use of cutting-edge data capture technology designed to ensure peace and reduce the physical and economic costs of conflict. CSO works diligently to enhance their capacity to address emerging conflicts. Big data is essential for these efforts.

[29] Ian Armas Foster. IRS to Utilize Big Data to Improve Returns. Datanami.com. April 15, 2013. http://www.datanami.com/datanami/2013-04-15/irs_to_use_big_data_to_improve_returns.html.

CSO worked with over 15 countries in its first year. Eighty percent of its efforts were focused on four priority countries: Syria, Kenya, Burma, and Honduras. In these countries, CSO analyzed large datasets to examine patterns derived from human behaviors, electronic signals, social media elements, and other forms of technological and non-technological data. Analysts used the data to pinpoint areas of potential violence or instability. CSO's analytic tools have been fairly accurate in predicting defection, unrest, and ally support. This knowledge has informed preemptive measures and resulted in saving lives, manpower, energy, and money.[30]

In Syria, CSO has invested $23 million in the country to strengthen their networks with the ability to communicate both internally and externally and build government capacity to support a transition. The money was also used to expand a network of nearly 500 Syrian activists, administrators, and journalists. The goal of these efforts was to provide insights into events inside Syria, expand assistance networks, and identify local leaders for action. CSO currently has a growing network of Syrian trainees and contacts working to monitor developments in critical areas inside the country.

In Kenya, more than 1,000 people died and 350,000 were displaced after allegations of corruption during the 2007 elections led to unrest and a political, economic, and humanitarian crisis. To prevent violence in future elections, CSO is supporting the Kenyans through the use of big data. The Kenyans have established an early-warning network in targeted hotspots to confront violence. CSO uses the Defense Advanced Research Projects Agency's (DARPA) Integrated Crisis Early Warning System, which aggregates more than 20 million news stories to identify trends in conflict at local, national, and regional levels, and then respond to potential threats. The CSO also aggregates data from sources such as networks, security, servers, databases, and applications from various parts of Kenya and around the world that offer the ability to consolidate monitored data and avoid missing critical events. This aggregation produces alerts at the moment abnormal activity appears.

CSO applies big data analytics to the developed world, allowing human and monetary resources to be conserved. Providing targeted assistance is essential to combating conflict in a growing digital age. CSO is not the only organization using big data: organizations like Ushahidi used big data to create a map that traced ethnic violence in real time during the Kenyan elections. Identifying patterns and trends in behaviors that lead to containing and preventing conflict is a powerful big data tool.

In State Government

Big data's growth is not limited to federal agencies. Increasingly, state and local governments are investing and experimenting with big data to improve decision-making, services, and efficiency. As noted by the National Association of State Chief Information Officers (NASCIO), big data is very important at the state level because it directly affects citizen outcomes.[31] Budgeting issues are of growing concern for state and local governments, and leaders face increasing pressure to make sound investments.

Interviews with CIOs in state government revealed that there were multiple big data projects underway. CIOs are working on various big data projects focused on getting a unified view of citizens and their interactions with public agencies, services, and goods. As an example, in most states, the data on citizens is spread across multiple databases (e.g., the Department of

[30] CSO: One-Year Progress Report. http://reliefweb.int/report/world/cso-one-year-progress-report.

[31] Is Big Data a Big Deal for State Governments? Nascio.org. August 2012. http://www.nascio.org/publications/documents/NASCIO_BigData_August2012.pdf.

Big Data R&D

In 2012, the Obama administration announced the Big Data Research and Development initiative to accelerate the pace of discovery in science while strengthening national security and transforming teaching and learning. The initiative provided more than $200 million to launch new big data projects across six federal departments and agencies.[32] President Obama noted that this initiative is an "all hands on deck" effort and challenged industry, research universities, and nonprofits to strive for new developments in big data. Innovative collaborations such as one between the National Science Foundation (NSF) and the National Institutes of Health (NIH) were established. The Core Techniques and Technologies for Advancing Big Data Science & Engineering competition was developed to advance core scientific and technological means of managing, analyzing, visualizing, and extracting useful information from large and diverse datasets. After the summer 2012 competition, NSF and NIH awarded 43 projects from universities across the nation. Funding was provided for several bold efforts, including a $10 million Expeditions in Computing project based at the University of California, Berkeley, that integrates three approaches for turning data into information. The foundation also provided grant support to "EarthCube," a system that allows geoscientists to access, analyze, and share information about the Earth, and a $2 million award to support training for undergraduates to use graphical and visualization techniques for complex data.[33] A selection of projects funded by the initiative includes:

- A $60 million annual investment by the Department of Defense (DOD) that will bring together sensing, perception, and decision support to make autonomous systems that can learn, maneuver, and make decisions on their own.
- A $25 million annual investment by the Defense Advanced Research Projects Agency (DARPA) to develop the XDATA program, which will develop computational techniques and software tools for analyzing large volumes of both semi-structured and unstructured data.
- The National Institutes of Health (NIH) and the National Science Foundation (NSF) are investing in Big Data science and engineering focused on managing, analyzing, visualizing, and extracting useful information from large data sets.
- A $25 million investment by the Department of Energy (DOE) to establish the Scalable Data Management, Analysis, and Visualization (SDAV) Institute. This Institute will bring together the expertise of six national laboratories and seven universities to develop new tools to help scientists manage and visualize data on the department's supercomputers.
- The U.S. Geological Survey (USGS) launched the Big Data for Earth System Science initiative to improve the understanding of species response to climate change, earthquakes, and ecological indicators.

Motor Vehicles, Department of Public Safety, and Department of Revenue). At the state level, getting a unified view of the citizen is challenging due to the disparate nature of the data, the structure of the databases, and the restrictions imposed by privacy and security policies. CIOs are working with their peers and elected officials to integrate databases, design information-sharing policies, and prototype new systems to enable agencies to better engage with citizens. A unified view of citizens can not only streamline costs of service delivery, but also improve a citizen's interactive experience with the public sector.

Massachusetts. In Massachusetts, Governor Deval Patrick made one such investment through the Massachusetts Big Data Initiative, which combines business and academic resources to

[32] Tom Kalil. Big Data is a Big Deal. Whitehouse.gov. March 29, 2012. http://www.whitehouse.gov/blog/2012/03/29/big-data-big-deal.

[33] Office of Science and Technology Policy, Executive Office of the President. Obama Administration Unveils "Big Data" Initiative: Announces $200 Million in New R&D Investments. Whitehouse.gov. March 29, 2012. http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf.

accelerate big data usage and research in Massachusetts through a number of projects. One such project is a $12.5 million, five-year partnership between Intel and MIT entitled "bigdata@CSAIL."[34] bigdata@CSAIL brings together leaders from government, business and academia to develop sophisticated techniques for capturing, processing, analyzing, storing, and sharing big data. Experts in hardware and software development, theoretical computer science, and computer security will coalesce to develop new architectures capable of sorting and storing massive quantities of information, as well as the algorithms that can process them. [35] bigdata@CSAIL's partners are looking to develop systems and platforms that are reusable and scalable across multiple application domains to solve society's challenges using data.

In Local Government

City governments are accumulating data at an accelerated rate, resulting in the commission of big data programs to address data management issues. Some city officials have found big data analytics to be useful in providing near-real-time information to the public through such powerful tools as environmental sensors and expeditious reporting of data on mobile devices. In addition, many cities are sharing raw data in the hopes that citizens will find uses for it. Large cities, including Chicago and New York City, offer their private citizens the use of open data to create public services in the form of mobile, web, or desktop applications (apps). Apps are now available in Chicago that show citizens which streets have been cleared after a snowfall, what time a bus or train will arrive, and how a request to fix potholes is progressing.

Cities are also starting to look at historical data and have made fascinating discoveries that impact the present.[36] For example, New York City data analysts noticed a trend in properties with tax liens levied against them. If a property had a tax lien, analysts found that there was a ninefold increase in the chance of a catastrophic fire on the property. Similarly, in Chicago, a call complaining about trash bins in certain areas is likely to be followed by a rat infestation the following week.

The United States is not alone in its pursuit of innovations to realize opportunities from mining large and complex datasets. The city of Dublin aims to improve traffic flow and mobility through a new program designed to identify and solve the root causes of traffic congestion in its public transportation network. Most major cities have begun investments in programs to make them "smarter."[37] Critical to a city's ability to become "smarter" is the capability to integrate datasets across heterogeneous systems, departments, and processes and analyze the data in real time to promote situational awareness and evidence-driven decision-making. CIOs are also working on leveraging data from geographical information systems (GIS) to enable better urban planning, transportation systems, and utility management.

New York City's Business Integrity Commission (BIC). BIC stands at the forefront of the city's vast business law enforcement infrastructure. BIC's central operation is the licensing of sanitation haulers and wholesalers. In an effort to preserve and maintain a level playing field among the city's wholesale market industries and private sanitation, BIC leverages expansive amounts of intelligence and information on multiple forms of corruption and organized crime within both industries. Tasked with regulating 2,000 businesses in New York City, BIC manages its

[34] Governor Patrick Announces New Initiative to Strengthen Massachusetts' Position as a World Leader in Big Data. Mass.gov. May 30, 2012.

[35] Larry Hardesty. MIT, Intel Unveil New Initiatives Addressing "Big Data." MIT.edu. May 31, 2012. http://web.mit.edu/newsoffice/2012/big-data-csail-intel-center-0531.html.

[36] City of Chicago—Chicago Digital. http://digital.cityofchicago.org/index.php/open-data-applications/

[37] Desouza, K.C. Designing and Planning for Smart(er) Cities. *Practicing Planner.* Winter 2012.

licensing operations, auditing functions, and discovery of corrupt dealings with data-driven processes.[38]

BIC uncovered serious sanitation issues that allowed illegal haulers to profit while licensed haulers lost money. New York City was experiencing criminal collection of commercial waste, particularly of recyclables such as paper, scrap metal, and yellow grease. Yellow grease is an inedible fat derived from cooking oil in the fast food industry. It is typically used to feed livestock, and to make soap, makeup, clothes, rubber, detergents, and biodiesel fuel.

Legally, municipal solid waste was collected and hauled to a permitted transfer station and the haulers either charged for their services or forwent a hauling service fee to recoup profits at the dumping site. Due to the increased value of recyclables, unlicensed individuals were criminally collecting waste and selling it. Recyclables such as yellow grease were becoming a growing commodity that many companies were converting into biodiesel fuel. In 2008, the Emergency Economic Stabilization Act of 2008 was passed into law. Widely known for bailing out the U.S. financial industry, this law also included significant implications for the biodiesel industry. The law extended the biodiesel tax credit and qualified all biodiesel for a $1.00 per gallon tax credit, including biodiesel made from yellow grease.[39]

To combat these unwanted practices, in 2012 BIC and the Mayor's Office of Analytics joined forces to spot perpetrators of this largely invisible crime. Using hotspot analysis, BIC and the Mayor's Office of Analytics cross-referenced industry data on grease production derived from the Department of Health and Mental Hygiene (DOHMH) restaurant permit data and the Department of Environmental Protection's (DEP) sewer backup data to target illegal activity. A Yellow Grease heat map was developed to visualize potential hotspots of unlicensed activity and direct enforcement efforts.

Since launching this effort, BIC reported achieving an increase in discovered violations by 30 percent and achieved a 60 percent reduction in manpower dedicated to grease enforcement.[40] BIC's efforts to lessen unlicensed, criminal waste hauling have had a number of positive effects. New information-sharing relationships with the DOHMH and DEP have been strengthened. Licensed haulers are no longer burdened with lost income due to criminal activity. Lastly, BIC is growing with the times and is allowing data to thoughtfully and effectively drive enforcement.

[38] New York City Business Integrity Commission. Annual Report 2012. http://www.nyc.gov/html/bic/downloads/pdf/pr/BICAnnualReport2012.pdf.

[39] Erin Voegele. U.S. Congress extends biodiesel tax credit. Biodieselmagazine.com. September 16, 2008. http://www.biodieselmagazine.com/articles/2825/u.s.-congress-extends-biodiesel-tax-credit/.

[40] Shari Hyman. Enforcement and Data: One New York City Agency's Vision for a Level Playing Field. Data-Smart City Solutions. August 13, 2012. http://datasmart.ash.harvard.edu/news/article/enforcement-and-data-280.

Findings from Interviews with Chief Information Officers

Study Findings

Finding One: Public agencies are in the early days of their big data efforts.

CIOs overwhelmingly report that they are just getting started with big data efforts. While big data as a concept has been discussed in the popular press and the academic literature for years, public agencies have not yet fully embraced the concept. CIOs have struggled with identifying a tangible pathway towards getting started on big data. While they see the value in synthesizing databases and using analytics, the challenge of big data has overwhelmed them. As one CIO notes:

> Big data is just too big for us. Where … big data begin[s] and end[s] is not known. I have been struggling to identify digestible bites for us to take to move on big data … But, if we are not doing something *big* is it big data or something else.

Many CIOs lack a tangible framework to guide their big data efforts. The lack of knowledge or examples of actionable methods for initiating big data projects poses a significant challenge. The financial, policy, and pragmatic intricacies required to set up a big data project are sometimes beyond the time constraints and expertise of many CIOs and their staff. In addition, the distinction between big data and conventional data is often not clearly understood. As one CIO tells us:

> When I think of big data, unstructured data comes to mind. Isn't that a critical element of big data? If so, then we are not doing anything in the big data space as we have not touched unstructured data. All of our data has some structure, and most of it is highly structured.

All CIOs are aware of the value proposition of data analytics for their agency. While they do not doubt that big data could revolutionize service delivery and streamline business operations, CIOs believe they have "limited bandwidth" to invest in these efforts. CIOs are also aware that at the present time, much of the existing data in their systems goes unanalyzed. CIOs know they need to do more with their existing datasets beyond transaction processing. Data is most often collected as an outcome of transactions and is only rarely used beyond the narrow scope of preparing reports on the transactions. CIOs are interested in broadening the way this data is used so that it can be a critical asset for strategic decision-making. As one CIO comments:

> We need to focus on analyzing the data we currently have stored [in our systems]. My guess is that we only analyze about 30% of it … there is a huge opportunity for us to work on the rest [of the data] and create value …"

About the Interviews

The research methodology was designed to focus on public agencies other than the most prominent ones in the realm of big data. Studying big data efforts at the Department of Energy or the National Security Agency, while important, may not have produced findings applicable to all public sector chief information officers (CIOs). CIOs at all three levels of government were interviewed:

- Federal
- State
- Local (city or county)

Before interviewing CIOs, we examined the secondary literature (e.g., news releases, articles, blog posts, etc.) to study an agency's experience with innovative information technology projects. We investigated their efforts in open data, social media, cloud computing, and big data with the goal of building a representative sample of agencies. We studied agencies that had a great deal of experience with deploying innovative technologies and those with limited or no experience.

In all, we interviewed 28 CIOs (six federal, six state, 16 city/county). We stopped interviewing when we had reached saturation in terms of discovery of new concepts and information. (After the completion of 22 interviews, key findings were compiled in a bullet point list format. The six successive interviews revealed no new information.) Nineteen of the 28 interviews were conducted by phone, with the rest conducted in person. With interviewee consent, interviews were recorded and extensive notes were prepared. After each interview, we prepared a brief summary statement of the major findings.

The major findings are grouped thematically on various concepts dealing with the management of big data programs. Theoretical concepts are organized into two major frameworks. The first framework segments findings based on the lifecycle of big data projects—initiation, conceptualization and design, execution, implementation, and post-implementation. The second framework classifies findings based on the key area they are focused on—organization, personnel, business processes, technology, and external environment.

Six CIOs who had not been interviewed were invited to comment on the report. Draft findings were presented at two management forums. The feedback received was overwhelming positive. All six of the CIOs who had not been interviewed confirmed that the findings resonated with their experience in managing big data programs. Only minor revisions were made to the overall report based on the feedback received.

Finding Two: Many CIOs fight the perception that big data is a passing fad.

CIOs say they have had to fight the perception that big data is just a buzzword or a fad. All CIOs note they have to tread carefully when discussing big data efforts, and a large proportion even note that they have avoided using the term "big data" because of its negative connotation in their agency. CIOs have had to spend significant efforts to educate their peer executives and agency heads about the value proposition for big data. They have to ensure that big data does not become viewed as merely a technology investment.

> Big data is a poorly defined term that has caused much angst ... Outside of IT no one gets the term …The few [managers] that do have some views on it, have been put off due to the privacy concerns because of the NSA's surveillance program … I explicitly told the IT group not to use the term in conversations with our customers across local government.

CIOs who had launched big data efforts received support by focusing on the organizational and business value that would be gained from investments in data analytics. Technologies that would be used for data analysis were placed in the background and were never the focus of the conversation.

Finding Three: Most CIOs are now primarily dealing with the issue of managing large volumes of data, integrating data across database systems, and building an analytical capacity to mine data.

All of the CIOs interviewed are now dealing with structured data. None of the CIOs interviewed have begun to explore the intricacies of managing unstructured data. While most CIOs realize that social networks and social media could provide them with rich data, they have not yet begun to tap into these spaces as part of their big data efforts.

Even when it comes to structured data, most CIOs admit that they are not dealing with datasets that are truly big (compared with datasets in the physical sciences). We did speak to CIOs who were truly managing large datasets (e.g., datasets at the Veterans Administration or Social Security Administration. These datasets are large, even if they are not so large as the ones receiving more attention in the literature). CIOs are experimenting with creating executive dashboards and using information visualization techniques to make the large volumes of information easily accessible and comprehensible.

Finding Four: CIOs report that some big data projects are now focused on streamlining business processes.

Some of the CIOs interviewed report that they are now focusing on identifying opportunities to streamline and reengineer their business processes through investments in data management. These CIOs are focusing on process management as a vehicle to gain support from stakeholders, both internal and external, to modify and bolster practices around how data is captured, stored, shared, analyzed, and applied.

Framing the use of big data as one of process management with opportunities for increased efficiency and effectiveness enables stakeholders to center their attention on the value of data management to the organization. As one CIO tells us, “A focus exclusively on data issues will not get people to care, but if you can get their attention on process improvement and innovation for organizational value then they do care…” CIOs understand that the early stage of any big data project is not glamorous. A lot of time and effort is needed to do data cleansing, data discovery, and even data validation in the early days of a big data initiative. These tasks are not necessarily ones that stakeholders will want to invest their resources on, unless they are convinced that they are going to gain significantly in terms of attaining their near-term and long-term objectives.

A CIO in a major U.S. city was able to improve business processes by making simple alterations to how data was captured by law enforcement officials. Data on suspects and incidents was captured in a free-form text field, which understandably led to inconsistencies in reporting as officers entered data in unique ways. For example, there were several dozen ways for data to be entered about a person’s eye color or hair color that resulted in significant delays in processing. The CIO convened an interagency group to help law enforcement address inefficient and ineffective business processes. Opportunities for significant advancement using better information sharing and data governance were addressed. A focus on business processes demonstrated the value of the data governance. By changing how data was captured from free-form textboxes to drop-down lists, the CIO was able to assist law enforcement in improving the efficiency by which suspects were processed and cases were referred to a judge for a hearing the next day.

Finding Five: CIOs do not anticipate significant investments in technology. Instead, they are focusing on more economical, strategic ways of using their current information technology assets.

Eighty percent of CIOs interviewed report that they would not need to make any significant investments into technology, both hardware and software, during the first few years of their big data programs. Instead they would need to find more economical and strategic ways of deploying their current information technology assets.

A good example of this trend is one CIO who is now working on a big data project in local government. During data integration and cleansing efforts, a precursor to launching an analytics program, he discovered that the agency's hardware was underused. Through a process of consolidation and thoughtful redeployment of data storage and applications, he gained the necessary resources for the big data project. In addition, when it came to software, most of the efforts revolved around customizing the existing applications and programs. Customizing existing applications and programs can include, for example, changing fields in one database from free-form to a drop-down list to ensure that captured data can be more easily analyzed. CIOs do say that as their big data programs mature and the complexity of their datasets increase, they will need to invest in both hardware and software.

Finding Six: CIOs report a need to bolster their human capital, including their analytical capability.

CIOs report that they are in the midst of bolstering their human capital through hiring new staff and training existing staff. In almost all cases, the stagnant, or diminishing, budgets in recent years have severely impacted the capabilities of most public sector IT units. CIOs report losing talented individuals to the private sector or early retirements. In addition, the inability to send staff for training and skill development has resulted in playing catch-up when it comes to big data skills.

> I do not have a single staff member who knows how to run simple regressions … Our staff has not played around with [Microsoft] Excel in any serious way … I am sure that this is not unique to our IT staff … Now when I hear about analytics being about networks, data mining, visualizations … we can be the installers of tools rather easily, but then who is going to train employees [outside of IT]."

In the federal government, current funding concerns caused by events such as the sequestration and shutdown have limited an agency's ability to take the long view in terms of planning for their human capabilities and needs. As a result, many are just trying to deal with crises and focus on the problems of today, which leaves them unprepared to tackle the opportunities of the future. Attracting talent capable of managing big data projects is a challenge for CIOs. Individuals equipped with the necessary skills to manage large-scale big data projects face strong incentives to work in the private sector. The private sector is an alluring career prospect because there is less bureaucracy and more money to be made. One CIO says:

> It is best to get talented individuals during an economic downturn like in 2009 when there were fewer jobs available. It is hard to attract people when budgets and resources are low.

CIOs report that they are now grappling with two primary staffing issues:

- Equipping and training the IT department to be proficient in analytics
- Increasing the analytical quotient of the entire organization

Most public sector organizations have little or no data analytics capabilities. There are some public sector entities that have analytical capacities (e.g., analysts working in the Department of Revenue). However, most IT employees have limited training to work with analytics. Public sector employees are more likely to be consumers of analytical reports rather than the generators of these products. CIOs remark that the various functional and business units of the organizations should be making investments in analytical capabilities. As one CIO tells us:

> We should be in the business of data governance and data policing … our business units should be the ones in charge of analytics. Relying on IT to be the producers of analytical reports and help with analytics is a sure way for us to upset everyone by being a bottleneck. We do not have the necessary expertise on the data to be able to do any meaningful analytics as well … business units are the experts and know what reports or data associations are going to be meaningful and valuable to them.

The lack of analytical capability across agencies is described as a severe deterrent to leveraging big data capabilities. CIOs note that there is a need to cooperate more closely with universities to ensure that the desired skills and training are made part of degree programs, especially graduate programs. As a CIO describes it:

> A few months back I volunteered to serve on an advisory board for a local university that was looking at creating a graduate program in data science. While they were trying to sell me on the idea and ensuring time and resource commitments … they did not realize that I was not doing this for altruistic reasons … we need students with skills in data analytics.

Finding Seven: CIOs are now exploring approaches to data governance.

The issue of data governance has been the primary concern for CIOs. All CIOs note that most of the data residing in their information systems is not readily suitable for analysis. Most of the data lacks integrity and could not be easily integrated across systems due to a lack of standardization in data definitions, and even if data could be integrated there are security and privacy considerations that need to be worked through. One CIO notes that data format varies greatly from discipline or career field; e.g., a geographer may maintain datasets with data fields that cannot be easily interpreted by others outside of their field.

CIOs say that poor data governance is the most critical factor holding up agencies in their efforts to pursue big data. Data governance shortcomings come in many forms:

- **Much of the data collected and stored in an agency's transaction processing systems lacks adequate integrity**. For example, in many agencies data that should be captured in structured formats (e.g., dropdown lists) is actually entered in free-form (textboxes). In addition, most agencies draw their source data from multiple systems, many of which are out of their direct control. As such agencies are at the mercy of third parties to ensure that the data being captured meets integrity standards.
- **Data governance is difficult when it is localized to individual systems and departments within an agency**. Each system and the data being captured is governed independently of the existing information system infrastructure. As such it is challenging to port data between systems and even reconcile data across systems during integration. This shortcoming prevents agencies from creating truly large datasets.
- **When one considers aggregating and connecting data across agencies, there is limited guidance in terms of policy and legal frameworks.** Because of this, agencies usually default to sharing information and data on a need-to-know basis rather than seeing ways to enrich their databases with more data. Privacy and security regulations, some CIOs have

argued, have swung too far in terms of promoting an environment where agencies do not have any incentives (and actually face severe hardships) for sharing data.

- **When it comes to data governance, agencies are already resource-constrained and thus do not have the "bandwidth" to invest in building better governance processes.** As one CIO notes:

> Ask anyone, our back-end systems and processes need TLC (tender loving care) … we have known this for ages ... still we continue to build on a fragile foundation … data governance is not sexy and no one wants to do it, yet it is our Achilles heel."

Finding Eight: CIOs do not recommend IT units as owners of big data projects.

CIOs caution against IT units being the owners or instigators of big data projects. Instead, they believe that senior management support is necessary for success. While senior management support may be a requirement for most IT projects, CIOs discuss several reasons why *active* senior management involvement is absolutely essential for big data projects:

- **Big data projects are transformational and involve multiple departments and agencies.** The IT unit of the agency is simply not going to have the organizational clout necessary to bring various stakeholders together.
- **The term big data has been bantered around in the press and for the most part has received a negative connotation in the public sphere.** It is absolutely essential for the senior executives of the agency, the mayor, the governor, and other significant stakeholders to be aware of a big data project and stay abreast of its progress.
- **Big data projects are messy and unglamorous at the start.** Agencies will need to invest time in the back-end tasks of cleaning up data, integrating business processes to capture data more efficiently, and designing collaborative data-sharing agreements. These tasks require agencies to invest in non-current operations that take them away from meeting their near-term goals. Moreover, these tasks are often unpleasant as they require one to dig deep into data, the history surrounding how data is captured and why, and then propose changes to the current way of doing things. Several CIOs note that some of these tasks become very political and it is imperative that "someone has your back." Therefore, it is critical that senior managers champion and advocate for the project in order to address the understandable organizational inertia that the project will face.

Finding Nine: CIOs believe that collaborative leadership is crucial for the success of big data projects and recommend the creation of working groups to oversee projects.

Interviewee responses indicate that collaborative leadership is crucial for big data projects. CIOs report creating interdepartmental or interagency working groups for big data projects. These working groups brought together key organizational stakeholders to move the project ahead. In organizations that were just beginning their initial foray into big data projects, these working groups were charged with writing a white paper (i.e., a briefing/position paper) to frame the value of big data for the agency. The white paper is meant to serve as an educational tool for senior managers and for awareness building among employees. Written documents are also designed to change perceptions of big data from an abstract concept to a legitimate method of building value for constituents.

In organizations that are in the midst of the initial stages of big data projects, working groups can serve multiple purposes:

- Being a clearinghouse for developing data governance standards and processes at the strategic level

- Representing their individual units and teams by being their spokesperson for the effort
- Maneuvering local resources necessary to advance the project

The working groups described were often chaired by a senior executive—e.g., a commissioner (city level) or a director/agency head (state level). An interagency working group comprised of senior-level executives and/or their designees was supported by operational interagency taskforces to work on the necessary details.

CIOs also rely heavily on their professional networks as they traverse the unchartered waters of big data. Most CIOs report checking in with their peers in other agencies to seek out information and insight. Based on our limited sample of interviews, we did find that CIOs who are more connected to their professional communities (e.g., they speak at industry conferences, are named by other interviewees as standout exemplars, etc.) are further along in their big data efforts. Several of these CIOs have led regional forums that bring their peers together to share best practices, solve problems collaboratively, and even advance big data projects at a regional scale.

Finding Ten: CIOs are becoming champions of analytics and evidence-driven decision-making.

CIOs note that they are becoming the de facto champions for analytics and evidence-driven decision-making within their organizations. This is a role that many of the CIOs did not envision for themselves, but one they are getting increasingly comfortable with. Given that most public agencies have been influenced by the trends of open government, CIOs have had to become comfortable with being stewards and disseminators of data. The dissemination, or opening up of data, has led to an increase in citizen apps and citizen-driven analytics. Given these trends, CIOs have now become champions for increasing the analytical capacity of their agencies and driving decisions based on evidence. As one CIO cautiously remarks:

> CIOs have a new calling and it is to be champions of analytics … If we are to justify our value it is going to be through analytics. System development and maintenance can be outsourced for the most part. Open data plus crowdsourcing has forced us to ask tough questions. If citizens can analyze data and build apps to benefit their communities and citizens then what are we doing … We [IT departments] can no more just claim that things are difficult or take time. If citizens can outdo us in analytics, how can we justify our budgets and expenses beyond just being a cost center? We need to innovate at the pace of citizens. Our future is tied to making our organizations information-centric.

Key Steps in Implementing a Big Data Project

Recognizing that big data projects are complex efforts to undertake, the following are best practices for public managers to consider as they make these efforts. These key steps, deduced through synthesizing critical success factors for big data projects as noted by CIOs, can form the three stages of a big data project.

Stage One: Planning

Public sector CIOs are familiar with the promise and perils of big data. The planning phase includes conceptualization of the project, which is vital for establishing a platform for success and ensuring that stakeholders are properly informed. This is an opportunity to lay the foundation of a quality project.

Various options must be evaluated, roles within a team must be agreed on, budgets must be secured, and other resources must be mobilized. Big data deployment can require large infrastructure changes, program designs, and agreements across agencies and departments. Attention to detail is important for success in this stage.

Step One: Before undertaking a big data project, CIOs must do their homework, which includes understanding relevant business and legal policies.

CIOs must understand the big data space and communicate intelligently about it. They must go beyond the hype and seek out examples of big data projects currently underway in other agencies. CIOs must ensure that they understand the value that a big data project can bring to an organization. In addition, understanding the challenges will assist CIOs as they begin to shape the effort and confront the project risks.

An understanding of the challenges and risks will help CIOs set realistic expectations and ward off the inevitable doubt and panic that will arise. CIOs that do not have realistic expectations about big data projects are likely to over-hype the promise and benefits while underplaying the risks and challenges—thereby setting themselves up for failure. One CIO notes that many are interested in big data because it is “trendy” and they tend to “jump into it” without knowledge of their own needs and capabilities. Dig into examples and look at what has worked and what has not and even contact individuals who have been featured in press stories. If CIOs do not have time to do their homework on big data, they should probably not commission a big data project.

The success of a big data project is dependent on the governance and policies in place around data, processes, and systems. CIOs need to spend time thinking through these complexities and the amount of work they are willing to invest in modifying and updating them for a big data effort. Data security and information access policies are just a few regulations put in place to protect public agencies and constituents. Many agencies have antiquated policies that do not account for current workplace technologies. Outdated policies should be revisited and adequately managed to make sure that big data projects are not halted due to noncompliance. Additionally, understanding the regulations that define the activities of potential partner agencies is crucial. Policies articulate the scope and manner in which services are to be rendered; upfront knowledge of possible barriers or similarities allows for better collaboration.

One CIO we interviewed notes that, when defining new opportunities, “it is important to understand the policies of other agencies, to identify shared constituents and minimize duplication of efforts.” Policies and legal frameworks play a critical role in advancing (or curtailing) a big data effort as they determine how data elements can be managed, accessed, and analyzed. Moreover, in cases where CIOs are considering an interagency big data effort, these documents help establish the rules for collaboration and even the expectations of each party.

Step Two: Before undertaking a big data project, CIOs must build a coalition that includes reaching out to peers.

Before CIOs embark on a big data project, they should seek out the experience of others. CIOs can gain immense knowledge from a simple conversation with their peers at other agencies. Professional networks are critical for CIOs to receive real-time information on the state of big data projects across the public sector. As one CIO notes, “the failures are never discussed in the articles and white papers on big data … everything is painted as the next best thing since sliced bread. When I talk with peers, I get the real deal on [big data] projects.” Professional networks can extend beyond the IT community. CIOs have turned to academic institutions, think tanks, and the private sector to seek out information and lessons learned.

Big data projects require collaborative leadership (Finding Eight). CIOs cannot take on this challenge on their own. Even if they could, they should not. Coalitions can be structured as working groups or advisory groups. These groups should extend the CIO’s reach and influence in the organization while at the same time enabling the concept of big data to be situated within the realities of the agency and/or the operating environment. Coalitions can go a long way in furthering agendas and creating inroads to new partnerships or information. CIOs will need to perfect their “elevator pitch” for big data to engage people in a coalition. The elevator pitch should explain how an investment in data management will allow the agency to tackle an existing problem more effectively and efficiently or take advantage of a new opportunity. As one CIO notes:

> I prepared a two-minute pitch that I could share with my peers … distill why should your peers care about big data or any IT program and sell it to them in a language they understand.

Step Three: In communicating big data projects, CIOs must define the broader opportunity.

Choosing the right opportunity for a big data project is critical. When choosing an opportunity, it is important to focus on how the project creates value for the citizens. As one CIO notes, “We want to choose a project that the mayor can highlight in his State of the City or one that the governor can brag about in the State of the State Address.” Choosing projects that benefit citizens and stakeholders directly is an effective way to draw attention and criticality to the project. In addition, it can be a way to contextualize how investments in big data can transform agency operations.

In addition, when choosing the opportunity, CIOs should focus on a data-driven transformative problem. At this stage of the process, CIOs should focus on the broad opportunity that exists for a big data project. In the next stage, CIOs should get to the specifics. Keeping the opportunity broad at the start allows CIOs more flexibility to engage other stakeholders and give them an opportunity to shape the program. A common strategy employed by CIOs is to outline the broad opportunity in the form of a working paper or position paper. This paper looks at the opportunities that exist within an agency for superior data management. The working paper then becomes the platform for having strategic discussions and deliberations.

Step Four: In selecting a big data project, CIOs should begin with the lowest-hanging fruit.

In choosing where to get started, CIOs overwhelmingly note that it is best to begin with the easiest opportunities first. If CIOs can begin a big data project by tackling data that is "public," they should do that before they get involved with "private" data. If CIOs can modernize existing technologies and processes to manage data more effectively and efficiently, they should do that before trying to create new processes.

CIOs that have witnessed success with their big data efforts note that they began by addressing problems that were simple, yet were visible pain points for an agency. Choosing the visible pain points and building a data-driven solution helps win support for the overall program. During the planning phase, it is advisable to build maps of the data elements and their interconnections. These maps are valuable tools to uncover intricacies such as data dependencies, interactions among data elements, and even organizational and political elements (e.g., who owns a given data element and governs it). Maps also help CIOs and their staff visualize the data space and identify areas for interventions. For example, an opportunity (e.g., lower transaction costs or lower rate of errors) may arise if the capturing of a given data element is streamlined across all systems. Maps also help in discovering areas where the organization may need to collect or integrate external data feeds into those that are presently not being captured. A big data project can help increase the situational awareness of analysts and policy makers on particular issues by providing them with richer pictures of the situation.

Step Five: In starting a big data project, CIOs must ensure strategic alignment of the project. This includes lining up sponsors.

Big data projects do not operate in isolation to other IT projects underway at any agency. Hence, alignment between projects becomes a critical concern for CIOs. CIOs report that, unless there is strong alignment between the big data projects and other efforts at the organization, there is a high chance that the big data project will fail due to it being perceived as:

- A distraction from core efforts
- A competing priority that is pulling away valuable resources

CIOs have tried to embed phases of a big data project into existing IT efforts. For instance, if there is already a project underway in an agency that is focused on the maintenance and upgrade of a given application, it might be easy to add two or more tasks centered on data cleaning and validation. Weaving data governance issues into every IT project helps streamline big data projects. This can be done by having a data management plan submitted for review to the data governance board before commissioning a project. CIOs can embed this effort into existing IT efforts to ensure ongoing project alignment. CIOs note that one might consider embedding a set of core principles into every IT project to tackle data governance issues so that work being done can be ready for a future big data effort with lower costs.

As discussed in Finding Seven, big data projects are unlikely to succeed without active support from senior management. These projects need a sponsor—someone who is willing to champion the project during moments of controversy or discomfort. It is important that someone with clout is willing to weather the proverbial storms that often accompany the initiation of big data efforts. Finding supporters that understand the need and benefits of big data and possess the credibility to properly communicate those benefits to others within and outside the organization is paramount.

Under almost no circumstances should an IT office conduct a big data project without a senior executive of the agency committed to sponsoring it. As noted, controversies around privacy and security are sure to arise. Having an executive sponsor behind the project will foster greater resilience in times of doubt or discomfort. Also, since big data projects often require large amounts of back-end work, an executive sponsor that supports the project and actually has the authority to make such a project happen is extremely important.

Step Six: In leading a big data project, CIOs should be privacy advocates.

Preserving and respecting the privacy of the public is of the utmost importance. While it goes without saying, most, if not all, public agency leaders are concerned with maintaining citizens' privacy. Privacy can be overlooked or undermined as an unintended consequence of big data projects. Government policies lag behind the growth of technology and the explosion of data. The Privacy Act of 1974 governs the collection, maintenance, use, and dissemination of personally identifiable information about individuals that is maintained in systems of records by federal agencies.[41] The Act was amended in 1988 to establish procedural safeguards regarding agency use of records when performing certain types of computer matching.[42] However, part of these laws may now be obsolete and require updating. It is anticipated that these laws will catch up eventually, but until they do, organizational leadership must take their role as privacy advocates seriously.

CIOs should be acutely aware of privacy and security considerations as discussions on data are taking place. This will be critical to project success. Ultimately, if CIOs are aware of these issues and advocate for care in their handling, this will be reflected positively in how the project proceeds and is perceived by stakeholders. Privacy and ethical considerations around data collection, integration, analysis, and dissemination should be discussed openly and sincerely. Seeking clarity from legal counsel is essential. In addition, speaking to the current stakeholders about their views on the implications for any changes to current data management practices from an ethical or privacy standpoint will lead to useful information for consideration.

Step Seven: Use taskforces in implementing a big data project.

A working group, or taskforce, should oversee the project and shape its direction. The working group should have the requisite expertise to oversee the project. Expertise should span both the technical and organizational dimensions. The ideal taskforce will bring forth complementary skills that cover the technology, business, and policy dimensions of the big data project:

- The technology dimension will include representatives from the IT unit.
- The business unit will have representatives that deal with executing tasks that generate or employ data being managed.
- The policy dimension will have representatives who can speak to the legal and governance restrictions that govern what can and cannot be done with data.

Each of these perspectives is valuable and must be included so as to ensure that the big data project does not run into any major surprises. One of the critical roles to assign to the taskforce

[41] The Privacy Act of 1974 (Pub.L. 93-579, 88 Stat. 1896, enacted December 31, 1974, 5 U.S.C. § 552a).

[42] The Computer Matching and Privacy Protection Act of 1988, (Pub.L. 100-503, amended the Privacy Act of 1974).

is that of the spokesperson. Ideally, there should be one individual to give regular updates to stakeholders and keep the senior sponsor apprised of any issues.

Step Eight: CIOs should outline expected resistance and plan for it when undertaking a big data project.

CIOs can safely assume that they will likely run into some resistance from various pockets of the organization when undertaking a big data project. In undertaking a big data project, CIOs are asking other parts of the organization to invest time on a project that may limit their ability to complete their core tasks. Other parts of the organization may also consider the project an IT fad, dismiss its value, and argue that it is a waste of resources. Additionally, big data projects often expose data that some may wish would stay buried.

One CIO interviewed for this study notes that his city's open access program caused internal strife because it gave city employees access to other city employees' information, resulting in discomfort throughout the organization. There will be political repercussions for analyzing data that was never looked at before. This is especially true if the big data project has anything to do with increasing efficiency of operations. Outlining the various sources and types of resistance upfront can help CIOs build an educated campaign and pitch for the project.

Step Nine: CIOs should develop key performance indicators before starting a big data project.

The development of key performance indicators around a big data project is critical. CIOs struggle to create usable key performance indicators. Development of such indicators, which focus on both *process* and *outcome* measures, is important for ensuring the success of the project:

- **Process measures** are traditionally centered on improving efficiency. These measures capture gains in process execution in terms of quicker completion times, lower costs of operations, etc., by advancing data governance systems.
- **Outcome measures** are focused on how customers perceive the service being delivered. These involve measures such as improved customer service, increased customer value, etc.

Key performance indicators should be communicated to the requisite stakeholders. Baseline data on organizational processes should be captured before the project begins. This will allow meaningful comparisons of outcomes, both before and after project commencement. Performance indicators should make sense to the business units involved and offer information on what the unit actually needs (not useless esoteric measures). Sharing statistics on how big a dataset is or how many reports are being generated might be useful to the IT department, but these metrics have limited value to the rest of the organization. Hence, it is critical that the chosen measures resonate with business functions, goals of the organization, and impact on citizens.

Step Ten: Design a risk mitigation plan before starting a big data project.

The power of big data analytics is great, as is the risk. Critical to the discussion around big data is how to leverage its power and maintain safety and privacy. Public sector databases are easy targets for hacking, fraud, identity theft, privacy breaches, and other misuses of data. All of the CIOs interviewed note a focus on privacy issues related to citizens' information and how it is accessed. A risk mitigation plan should be developed to assess the impact of compromised data, e.g., identifiable citizen information and enterprise data. Development of a risk mitigation plan is an essential component of the planning stage, and will help design processes to lessen the risks associated with compromised data. It is important to consider who has access to data, how much sensitive information is returned when database queries are made, and what the physical security surrounding server rooms is.

Understanding the risk and planning for it are extremely important to the discussion of big data and data stewardship. In addition to a risk management plan, there should also be a communications plan developed. The communications plan should include dealing with the press, academia, other agencies, etc. Communications on the big data program should be carefully managed and to a large degree controlled. One of the key reasons for asserting control on the communications is to ensure that the message being sent out is:

- Accurate (i.e., reflective of the goals and activities of the program)
- Goal-oriented (i.e., advances the goal of the agency or the program)

As one CIO notes:

> Some of my peers love being in the spotlight ... I do not ... Until we have something tangible to share about outcomes, I prefer not talking about it [the big data project] ... worst case it will be misrepresented in the media, best case it will be overhyped and we will be set up for failure.

Stage Two: Execution

After the planning stage is complete, it is time to begin the project. Executing a big data project requires ongoing attention from the project's advisory group and the staff managing the project. Learning and establishing best practices for project management is important. Organizational proficiencies or inefficiencies can bring about the success or failure of the project.

Step Eleven: When a big data initiative is underway, CIOs should constantly gauge the pulse of the program.

It is important to continuously stay abreast of project status. The previously developed timeline should be referenced for quality control. Consistent monitoring of a big data project can preempt major problems and allow for the development of creative solutions. Creating a project dashboard of the key performance indicators and status elements will be important.

Most CIOs interviewed use formal or informal dashboards on projects and these can be customized to meet the needs of a big data effort. Traditionally, status issues or concerns are marked as red, yellow, or green:

- Red: an issue that has reached a critical stage and warrants immediate attention
- Yellow: an issue that has risen above a given acceptable level
- Green: an issue of all-clear and normal status

CIOs say they need to regularly check the pulse of the program both from a process and outcome perspective. In addition, they need to constantly gauge the conditions in the environment, especially in terms of any sentiment toward the project. Appropriate and timely communications, along with other interventions, can help address the issues and nip potential problems in the bud.

Step Twelve: During the big data initiative, it is crucial to communicate, communicate, and communicate.

Big data is a tool that CIOs can use to answer key questions and provide better services to citizens. Big data analytics makes the computational work that humans can't achieve on their own possible. Yet humans are vital during all stages of big data projects, and eventually use data products to inform decisions.

Communication about milestones, inefficiencies, successes, and failures will help an agency and peers gain a better understanding of big data. In instances where data is shared between agencies, constant coordination, communication, and feedback is necessary to ensure mission success. One CIO remarks:

> Within the city government, we are moving away from the silo mentality. Now, we see the city as a whole and not individual departments. We are making changes in the corporate culture and making technology more amenable to that.

More and more agencies are now recognizing the power of collective action and are taking steps to engage more for greater success.

Step Thirteen: Throughout a big data project, CIOs must manage scope creep.

Related to the termination step discussed below is the issue of managing scope creep. CIOs note the propensity for scope creep in the project. This especially occurs as stakeholders see the progress being made on the project and think about additional ways for the data to be integrated, analyzed, visualized, and mobilized within an agency.

It is critical that CIOs keep a watchful eye for scope creep and be clear on the boundaries of the current effort and how future revisions and additions will be made. One approach might be to take the model that Google follows and release products in *beta*. Given that the product is released as a *beta,* it does not have to be the final product or the one that contains all the bells and whistles. New ideas and suggestions can be captured during a project and worked into the next release or update. Failure to manage scope creep could lead to a situation where the CIO will have to continuously adjust the project plan and deliverables, which is not desirable.

Step Fourteen: During a big data project, CIOs must stay focused on the data and not get caught up or distracted by the technologies involved in the project.

As one CIO notes, "Information technology is less about the technology and more about information." It is necessary to understand the agency's needs and capabilities. Instead of purchasing all-new technology, agencies may only need to refit or repurpose existing software or hardware. It is easy to get enamored with "technology bling" (a term used by one interviewee) and lose sight of the project's goals, which should not be to get new fancy gadgets.

Multiple CIOs report that the minute their agency announces an effort on big data (or any other major data activity), they are bombarded with calls from sales consultants who inquire and try to sell products and services. Having a clear focus on the goal of the project, which is to leverage data and manage it more effectively toward a business outcome, helps keep everyone focused. As a guide, the discussion on technology and its various nuances should not dominate every team meeting, as the real challenge is going to be managing data and the various issues surrounding this. Once data management issues are resolved from an organizational and policy viewpoint, the technology part can begin and should be fairly straightforward.

Step Fifteen: If necessary, CIOs should pull the plug on a big data project.

Given the nature of big data projects, and especially the level of investment and attention they receive, it might be hard to deem a project unsuccessful. There is a tendency to avoid pulling the plug on the project and to continue focusing on the sunk costs. This escalation of commitment to a failing course of action can not only continue to worsen the situation, but also negatively impact the state of the entire IT department.

One strategy suggested by a CIO is to outline clearly at the project's beginning the conditions under which the project would be stopped. Thinking through these upfront not only helps in setting realistic expectations of the project but also will sensitize the team to look for signals of trouble and discuss them openly during the team meetings.

Stage Three: Post-Implementation

After a big data project has been implemented, ongoing attention is still needed but in a different way. The post-implementation phase is an opportunity for the agency to review the project, identify strengths and weaknesses, and plan for the next project.

Step Sixteen: At the end of a big data project, conduct a postmortem and an impact analysis.

Documentation is crucial throughout the entire project; lessons learned from all stages—from planning to execution—should be recorded to ensure past mistakes are not repeated in the future. These lessons will be integral to the organization's growth, as well as invaluable information for colleagues in the field. Issues that many CIOs experience, such as data standards, limited funds, and data cleaning, are reoccurring. Through connections made from coalitions and advisory groups, many of these successes and failures can be shared with peers. Big data analytics will consistently evolve as technologies and needs change. Navigating the challenges of big data will be noteworthy and important in conversations about moving forward with big data.

Most CIOs note that they are interested in and actively searching for examples of big data projects around the country. Insight into management of big data projects is paramount to ensure the success of future initiatives. One important element of conducting a postmortem is that it should not be used for *evaluation* or to point figures at individuals or events. Unless people are protected to share their true experiences and learning episodes, the postmortem exercise will not be of any value.[43]

Conducting a thorough impact analysis is critical for conveying the value of a big data project. As noted above, the development of key performance indicators and the collection of baseline data are important. Impact analyses need not only account for improvement in process measures, such as efficiency of resource usage, but should also be tied to organizational value measures. Organizational value measures can include gains in service delivery measures, cost reduction outcomes, and the ability to innovate and take on new opportunities. Once an impact analysis is done, publicize it.

Step Seventeen: Identify the next project.

In the post-implementation phase, CIOs will have an opportunity to identify the next big data effort. This effort will naturally originate from the lessons learned and the results of the first project. For CIOs that have embarked on their second project, they have used the first project to decide on:

- The practices and processes they are going to replicate for the next project
- Areas of opportunity that arose during the first project and can be tackled

[43] Kevin Desouza, Togeir Dingsøyr, and Yukika Awazu. Experiences with Conducting Project Postmortems: Reports vs. Stories. 2005. *Software Process Improvement and Practice.* Issue 10, Volume 2.

The CIOs caution against launching another major big data project immediately following completion of the first project. According to one CIO, "you have to be careful to manage fatigue … these [big data] projects take a lot out of the organization and a breather may be in order … do not immediately jump into another major effort." As noted by the CIO, the organization needs time to recoup energy after the big data project, and it is advisable for the IT department to give the organization some time to recover. One additional benefit of waiting before launching the next effort is that it gives CIOs more time to collect evidence on the performance and benefit of the first project. This information will help CIOs make a stronger case for the next project.

Conclusion

Big data is of increasing importance to public agencies. This research has determined that most government entities are just beginning to develop the analytical capabilities necessary to process big data. Big data programs are beginning to emerge and are expected to increase in prominence in the near future. Data analytics holds great promise for increasing the efficiency of operations, mitigating risks, and increasing citizen engagement and value. We expect public agencies to take measured and calculated approaches as they traverse big data opportunities.

To ensure the success of big data initiatives, it is important for the IT organization to work with their business executives to:

- Build a strong case for the value of a big data project
- Develop a collaborative leadership model where the business executives act as leads and sponsors of the big data effort
- Develop key performance indicators to gauge the success of the big data effort

CIOs need to take great care to develop inter-organizational taskforces to address issues of data governance and sharing. CIOs are better served tackling public data before undertaking a project that involves private or sensitive data where security concerns are present. Given the limited resources that CIOs have to invest in new projects, and the competing demand of managing current technologies and programs during the fiscal year, it is imperative for CIOs to find ways to embed big data projects into existing efforts. Securing visible successes early on is critical. Big data projects need to be scoped carefully so as to secure the necessary buy-in from internal and external stakeholders. Showing success early on can go a long way in building allies for a big data project.

Over the next few years, nearly all public agencies will be grappling with the issue of how to integrate their disparate data sources, build analytical capacities, and move toward a data-driven decision-making environment. While most of these efforts will be focused on managing structured data that originates from trustworthy sources and is generated at high frequencies, over time agencies will have to contend with the challenges of managing data of varied types, reliability, credibility, and focus—this is when we will be truly leveraging big data as it is presently defined.

Acknowledgements

I would like to thank all the CIOs who generously gave their time and energy to participate in this research project, both as interviewees and as reviewers of the report. Kendra Smith assisted in conducting literature reviews, interviews, and preparing of case studies. Kena Fedorschak assisted with editing of the report. All errors and omissions are solely my responsibility.

About the Author

Kevin C. Desouza serves as the Associate Dean for Research at the College of Public Programs and is an associate professor in the School of Public Affairs at Arizona State University. He is also serving as the Interim Director of ASU's Decision Theater. He is a Senior Faculty Research Associate at the Center for Organization Research and Design and a faculty affiliate at the Center for Policy Informatics.

Immediately prior to joining ASU, he directed the Metropolitan Institute in the College of Architecture and Urban Studies and served as an associate professor at the Center for Public Administration and Policy within the School of Public and International Affairs at Virginia Tech. From 2005 to 2011, he was on the faculty of the University of Washington (UW) Information School and held adjunct appointments in UW's College of Engineering and at the Daniel J. Evans School of Public Affairs. At UW, he co-founded and directed the Institute for Innovation in Information Management (I3M); founded the Institute for National Security Education and Research, an interdisciplinary, university-wide initiative, in August 2006 and served as its director until February 2008; and was an affiliate faculty member of the Center for American Politics and Public Policy.

Desouza has authored, co-authored, and/or edited nine books and has published more than 125 articles in prestigious practitioner and academic journals. Desouza's work spans multiple disciplines and his publications have appeared in leading information systems, management, public administration, innovation, technology management, and software engineering journals. His most recent book is *Intrapreneurship: Managing Ideas Within Your Organization* (University of Toronto Press, 2011). His work has also been featured in a number of publications such as *Sloan Management Review, Harvard Business Review, Businessweek, Washington Internet Daily, Computerworld, KM Review, Government Health IT*, and *Human Resource Management International Digest*. Desouza is the author of a 2012 report for the IBM Center for The Business of Government, *Challenge.gov: Using Competitions and Awards to Spur Innovation.*

Desouza has advised, briefed, and/or consulted for major international corporations, non-governmental organizations, and public agencies on strategic management issues ranging from management of information systems to knowledge management, competitive intelligence, and crisis management. He has also advised budding entrepreneurs launching new initiatives and organizations. Desouza's current research includes strategic management of information systems in the public sector, innovation within public agencies and multi-sectoral networks, design of smart cities, and information and knowledge sharing networks.

Key Contact Information

To contact the author:

Kevin C. Desouza
Associate Dean for Research, College of Public Programs
Associate Professor, School of Public Affairs
Arizona State University
Office of the Dean
Suite #750
411 N. Central Avenue, Mail Code 3520
Phoenix, AZ 85004-0685
(206) 859-0091

e-mail: kev.desouza@gmail.com
http://www.kevindesouza.net

Reports from IBM Center for The Business of Government

For a full listing of IBM Center publications, visit the Center's website at **www.businessofgovernment.org.**

Recent reports available on the website include:

Acquisition

Eight Actions to Improve Defense Acquisition by Jacques S. Gansler and William Lucyshyn

A Guide for Agency Leaders on Federal Acquisition: Major Challenges Facing Government by Trevor L. Brown

Controlling Federal Spending by Managing the Long Tail of Procurement by David C. Wyld

Collaborating Across Boundaries

Engaging Citizens in Co-Creation in Public Services: Lessons Learned and Best Practices by Satish Nambisan and Priya Nambisan

Coordinating for Results: Lessons from a Case Study of Interagency Coordination in Afghanistan by Andrea Strimling Yodsampa

Collaboration Between Government and Outreach Organizations: A Case Study of the Department of Veterans Affairs by Lael R. Keiser and Susan M. Miller

Using Crowdsourcing In Government by Daren C. Brabham

Developing Senior Executive Capabilities to Address National Priorities by Bruce T. Barkley, Sr.

Beyond Citizen Engagement: Involving the Public in Co-Delivering Government Services by P. K. Kannan and Ai-Mei Chang

Implementing Cross-Agency Collaboration: A Guide for Federal Managers by Jane Fountain

Improving Performance

Incident Reporting Systems: Lessons from the Federal Aviation Administration's Air Traffic Organization by Russell W. Mills

Predictive Policing: Preventing Crime with Data and Analytics by Jennifer Bachner

The New Federal Performance System: Implementing the GPRA Modernization Act by Donald Moynihan

The Costs of Budget Uncertainty: Analyzing the Impact of Late Appropriations by Philip G. Joyce

Using Technology

Cloudy with a Chance of Success: Contracting for the Cloud in Government by Shannon Howle Tufts and Meredith Leigh Weiss

Federal Ideation Programs: Challenges and Best Practices by Gwanhoo Lee

Rulemaking 2.0: Understanding and Getting Better Public Participation by Cynthia R. Farina and Mary J. Newhart

The Use of Data Visualization in Government by Genie Stowers

Mitigating Risks in the Application of Cloud Computing in Law Enforcement by Paul Wormeli

Challenge.gov: Using Competitions and Awards to Spur Innovation by Kevin C. Desouza

Working the Network: A Manager's Guide for Using Twitter in Government by Ines Mergel

IBM Center for
The Business of Government

About the IBM Center for The Business of Government

Through research stipends and events, the IBM Center for The Business of Government stimulates research and facilitates discussion of new approaches to improving the effectiveness of government at the federal, state, local, and international levels.

About IBM Global Business Services

With consultants and professional staff in more than 160 countries globally, IBM Global Business Services is the world's largest consulting services organization. IBM Global Business Services provides clients with business process and industry expertise, a deep understanding of technology solutions that address specific industry issues, and the ability to design, build, and run those solutions in a way that delivers bottom-line value. To learn more visit: ibm.com

For more information:
Daniel J. Chenok
Executive Director
IBM Center for The Business of Government
600 14th Street NW
Second Floor
Washington, DC 20005
202-551-9342
website: www.businessofgovernment.org
e-mail: businessofgovernment@us.ibm.com

Stay connected with the IBM Center on:

f YouTube

t in govloop

or, send us your name and
e-mail to receive our newsletters.