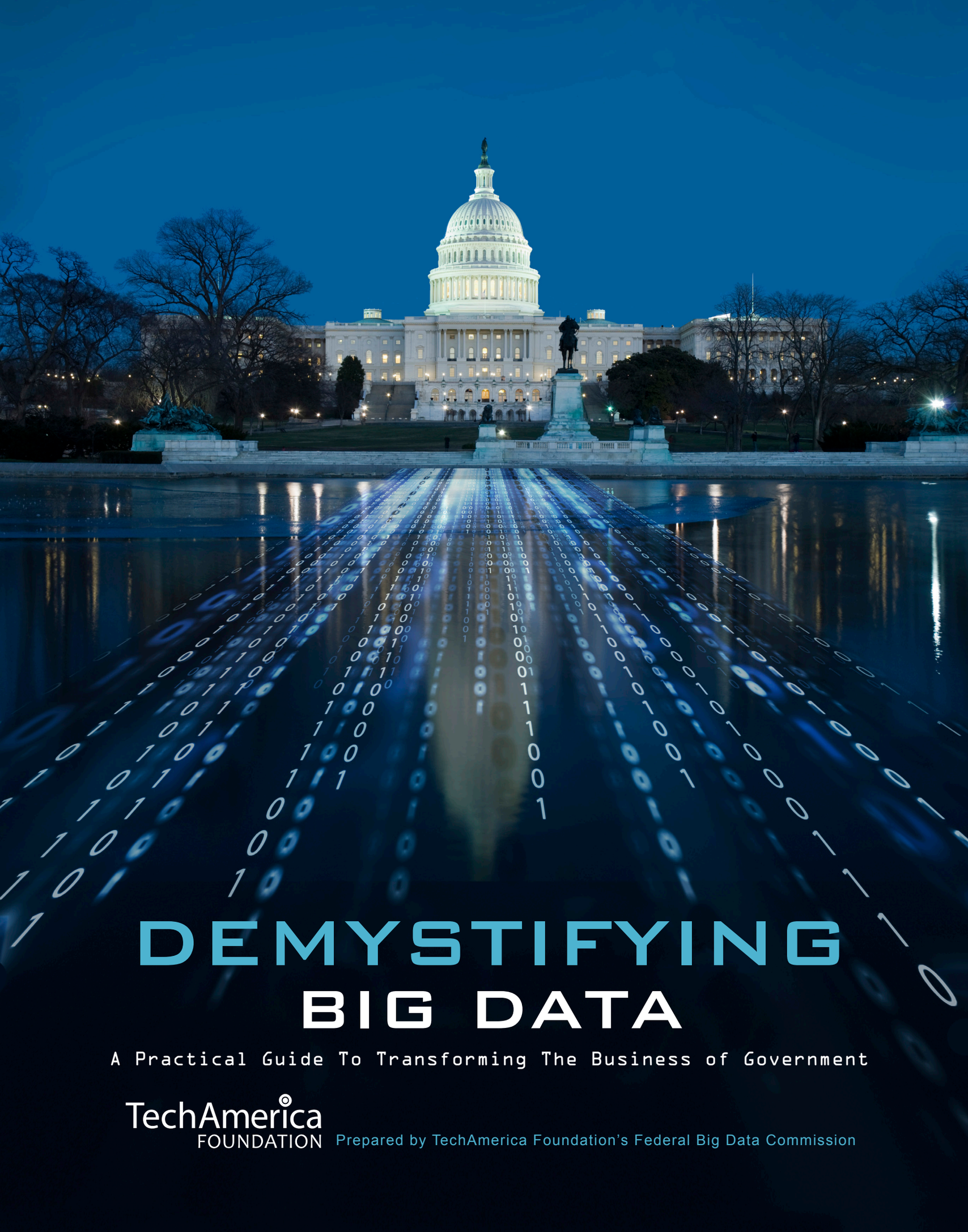


DEMYSTIFYING BIG DATA

A Practical Guide To Transforming The Business of Government

TechAmerica FOUNDATION

Prepared by TechAmerica Foundation's Federal Big Data Commission

Listing of Leadership and Commissioners

Leadership

Steve Mills (Co-Chair)
Senior Vice President and Group Executive
IBM

Steve Lucas (Co-Chair)
Global Executive Vice President and General Manager, Database and Technology
SAP

Leo Irakliotis (Academic Co-Chair)
Western Governors University

Michael Rappa (Academic Co-Chair)
North Carolina State University

Teresa Carlson (Vice Chair)
Vice President Global Public Sector
Amazon Web Services

Bill Perlowitz (Vice Chair)
Chief Technology Officer, Science, Technology and Engineering Group
Wyle

Commissioners

Manish Agarwal
Attain

Sevag Ajemian
Globanet

Tim Bradley
MicroStrategy

Rich Campbell
EMC Corporation

Stephen Coggeshall
ID Analytics

Bill Cull
Splunk

Eric Gillespie
Poplicus, Inc.

John Igoe
Dell

Prem Jadhwani
GTSI Corp.

Richard Johnson
Lockheed Martin IS&GS

Yogesh Khanna
CSC

Bilhar Mann
CA Technologies

Ray Muslimani
GCE

James Norton
General Dynamics
C4 Systems

Mike Olson
Cloudera

Steven Perkins
Grant Thornton LLP

Raghu Ramakrishnan
Microsoft

Rich Rosenthal
TASC

David Shacochis
Savvis, A CenturyLink Company

Hemant Sharma
CGI

Rick Sullivan
Hewlett-Packard

Michael Van Chau
MEI Technologies

Dale Wickizer
NetApp

Federal Big Data Commission Staff

Chris Wilson
Staff Director

TechAmerica Foundation Staff

Jennifer Kerber
President

Table of Contents

Table of Illustrations

Foreword

In recent years, federal, state and local government agencies have struggled to navigate the tidal wave of sheer volume, variety, and velocity of data that is created within their own enterprise and across the government ecosystem. As this tidal wave has swept across government, "Big Data" has arisen as the new ubiquitous term. Everyone is talking about Big Data, and how it will transform government, both in Washington and beyond the Beltway. Looking past the excitement, however, questions abound. What is Big Data? What capabilities are required to keep up? How do you use Big Data to make intelligent decisions? How will agencies effectively govern and secure huge volumes of information, while protecting privacy and civil liberties? Perhaps most importantly, what value will it really deliver to the government and the citizenry it serves?

In order to answer these questions and to provide guidance to our federal government's senior policy and decision makers, the TechAmerica Foundation Big Data Commission relied upon its diverse expertise and perspectives, input from government representatives, and previous reports. The Commission's mandate was to demystify the term "Big Data" by defining its characteristics, describe the key business outcomes it will serve, and provide a framework for policy discussion.

Although there clearly is intense focus on Big Data, there remains a great deal of confusion regarding what the term really means, and more importantly, the value it will provide to government agencies seeking to optimize service outcomes and innovate. This confusion may be due in part to the conversation being driven largely by the information technology community versus line of business community, and therefore centering primarily on technology. This report approaches Big Data from the perspective of the key mission imperatives government agencies must address, the challenges and the opportunities posed by the explosion in data, and the business and inherent value Big Data can provide. The report breaks the discussion down into five chapters:

1. Big Data Definition & Business/Mission Value
2. Big Data Case Studies
3. Technical Underpinnings
4. The Path Forward: Getting Started
5. Public Policy

The Commission based its findings on the practical experiences of those government leaders who have established early successes in leveraging Big Data, and the academics and industry leaders who have supported them. The intent is to ground report recommendations in these best practices and lessons learned, in an effort to cut through the hype, shorten the adoption curve, and provide a pragmatic road map for adoption.

Executive Summary & Key Findings

Big Data has the potential to transform government and society itself. Hidden in the immense volume, variety and velocity of data that is produced today is new information, facts, relationships, indicators and pointers, ***that either could not be practically discovered in the past, or simply did not exist before.*** This new information, effectively captured, managed, and analyzed, has the power to enhance profoundly the effectiveness of government. Imagine a world with an expanding population but a reduced strain on services and infrastructure; dramatically improved healthcare outcomes with greater efficiency and less investment; intensified threats to public safety and national borders, but greater levels of security; more frequent and intense weather events, but greater accuracy in prediction and management. Imagine a world with more cars, but less congestion; more insurance claims but less fraud; fewer natural resources, but more abundant and less expensive energy. The impact of Big Data has the potential to be as profound as the development of the Internet itself.

Harnessing Big Data also will serve the key objectives and recommendations described in the Digital Government Strategy report the White House released on 23 May 2012 – "Digital Government: Build a 21st Century Platform to Better Serve The American People" (Digital Government Strategy). A primary component of the Digital Government Strategy is to "unlock the power of government data to spur innovation across our nation and improve the quality of services for the American people." Big Data promises to fulfill the very essence of this objective.

The great paradox is that, as Big Data emerges as a new resource, we struggle to keep pace. We find it difficult to discover, understand, and leverage the information it contains, to find those true nuggets of knowledge that can improve the lives of everyday citizens and change the world. Although there is more data available, our ability to comprehend this data is reduced. The challenge lies in capturing the streams of Big Data that we need, effectively managing them, and extracting new and relevant insights.

The good news is that not only is it possible to extract value from Big Data, but the path is relatively straightforward. Leaders across government, academia, and private industry have made investments, have demonstrated success, and we now know what "good" looks like; there exist best practices from which we can define the path forward.

These experiences reveal that although the impact of Big Data will be transformational, the path to effectively harnessing it does not require government agencies to start from scratch with greenfield investment. Rather government can build iteratively on the capabilities and technologies it already has in place.

Perhaps as important, the path to harnessing the value of Big Data is now ***affordable***. It is this convergence of the availability of Big Data, the ability to harness it, and the affordability of doing so that brings government to an inflection point. The time to act is now.

Success in capturing the transformation lies in leveraging the skills and experiences of our business and mission leaders, rather than creating a universal Big Data architecture. It lies in understanding a specific agency's critical business imperatives and requirements, developing the right questions to ask, understanding the art of the possible, and taking initial steps focused on serving a set of clearly defined use cases. The experiences and value gained in these initial steps lead to more questions, more value, and an evolutionary expansion of Big Data capability that continually leverages prior investments.

It is instructive to remember the phenomenon of eBusiness. In the late 1990s, the buzz was that eBusiness was going to change the world. By 2005, the term largely had faded away, it became passé. Yet, looking across government and society, it is clear that effective organizations operate on the fundamental principles wholly consistent with the term. eBusiness did in fact change the world. One can argue that those organizations that successfully harnessed the power of eBusiness started with their operational challenges and requirements first, and asked, "How can the Internet help?," versus diving immediately into the technology. So it will be with Big Data. Successful government agencies will seek to define their requirements and use cases, and ask, "How can Big Data help?" versus setting out to deploy Big Data projects. Ten years from now, we may have forgotten the term, but its principles will underpin society.

Defining Big Data

• Big Data is a phenomenon defined by the rapid acceleration in the expanding volume of high velocity, complex, and diverse types of data. Big Data is often defined along three dimensions -- volume, velocity, and variety.

• This phenomenon represents both a ***challenge*** in making sense of the data available to governments, and an ***opportunity*** for government agencies that seek to exploit it to enhance the business of government.

• Addressing the challenge and capturing the opportunity requires advanced techniques and technologies to enable the capture, storage, distribution, management, and analysis of the information.

• Government leaders should strive to understand the "Art of the Possible" enabled by advances in techniques and technologies to manage and exploit Big Data. Example of use cases and live case studies are critical in understanding the potential of Big Data.

Leveraging Big Data – Early Experiences and Lessons Learned

• While Big Data is transformative, the journey towards becoming Big Data "capable" will be iterative and cyclical, versus revolutionary.

• Successful Big Data initiatives seem to start not with a discussion about technology, but rather with a burning business or mission requirement that government leaders are unable to address with traditional approaches.

• Successful Big Data initiatives commonly start with a specific and narrowly defined business or mission requirement, versus a plan to deploy a new and universal technical platform to support perceived future requirements. This implies not a "build it and they will come" transformative undertaking, but rather a "fit for purpose" approach.

• Successful initiatives seek to address the initial set of use cases by augmenting current IT investments, but do so with an eye to leveraging these investments for inevitable expansion to support far wider use cases in subsequent phases of deployment.

• Once an initial set of business requirements have been identified and defined, the leaders of successful initiatives then assess the technical requirements, identify gaps in their current capabilities, and then plan the investments to close those gaps.

• Successful initiatives tend to follow three "Patterns of Deployment" underpinned by the selection of one Big Data "entry point" that corresponds to one of the key characteristics of Big Data – volume, variety and velocity.

• After completing their initial deployments, government leaders typically expand to adjacent use cases, building out a more robust and unified set of core technical capabilities. These capabilities include the ability to analyze streaming data in real time, the use of Hadoop or Hadoop-like technologies to tap huge, distributed data sources, and the adoption of advanced data warehousing and data mining software.

Recommendations For Getting Started – For Agency Leaders

Big Data is a phenomenon characterized by the exponential expansion of raw data that is inundating government and society. It is already here and it is accelerating. The path to effective government described in the Digital Government Strategy lies in developing a set of capabilities to meet the challenge and harness its value. Perhaps unlike any other technical challenge the government now faces, Big Data will not be ignored.

The question lies in how to respond. The good news is that leveraging Big Data is affordable and early experiences offer best practices. As many government agency leaders take the first steps toward adopting Big Data solutions, the Commission makes the following recommendations:

1. Understand the "Art of the Possible" -- Explore the case studies contained in this report, posted on the TechAmerica Foundation Website, and otherwise in the public domain to find inspiration and practical examples.

2. Identify 2-4 key business or mission requirements that Big Data can address for your agency, and define and develop underpinning use cases that would create value for both the agency and the public.

3. Take inventory of your "data assets." Explore the data available both within the agency enterprise and across the government ecosystem within the context of the business requirements and the use cases.

4. Assess your current capabilities and architecture against what is required to support your goals, and select the deployment entry point that best fits your Big Data challenge – volume, variety or velocity. (The entry points are described in detail in the "The Path Forward: Getting Started" chapter of this report.)

5. Explore which data assets can be made open and available to the public to help spur innovation outside the agency. Consider leveraging programs like the Innovation Corps offered by the National Science Foundation, or the Start-Up America White House initiative.

Beyond the Individual Agency – Policy Considerations

From a policy perspective, the federal government should examine existing organizational and technical structures to find and remove barriers to greater Big Data uptake and, where needed, take action to accelerate its use. Specifically, the government should:

1. Expand the talent pool by creating a formal career track for line of business and IT managers and establish a leadership academy to provide Big Data and related training and certification.

2. Leverage the data science talent by establishing and expanding "college-to-government service" internship programs focused specifically on analytics and the use of Big Data.

3. Establish a broader and more long-lasting coalition between industry, academic centers, and professional societies to articulate and maintain professional and competency standards for the field of Big Data.

4. Expand the Office of Science and Technology Policy (OSTP) national research and development strategy for Big Data to encourage further research into new techniques and tools, and explore the application of those tools to important problems across varied research domains.

5. Provide further guidance and greater collaboration with industry and stakeholders on applying the privacy and data protection practices already in place to current technology and cultural realities.

Defining Big Data & Business/Mission Value

Big Data is not a technology, but rather a phenomenon resulting from the vast amount of raw information generated across society, and collected by commercial and government organizations. This phenomenon represents both a challenge in harnessing this volume of data, and an opportunity for government agencies who seek to enhance their effectiveness. This section describes the accelerating explosion of Big Data, the definition of Big Data, the characteristics of Big Data, the mission and business value Big Data promises, and potential use cases.

Current Trends and Explosion of Big Data

In recent years, federal, state, and local governments have come to face a tidal wave of change as a result of the drastic increase in the sheer volume, variety and velocity of data within their own enterprise and across the government ecosystem. For example, in 2011, 1.8 zetabytes of information were created globally, and that amount is expected to double every year. This volume of data is the equivalent of 200 billion, 2-hour HD movies, which one person could watch for 47 million years straight. The impact of this phenomenon to business and government is immediate and inescapable.

Because of the Internet and influx of information from multiple sources embedded within every fabric of our government, agencies will continue to struggle with managing large streams of data. Our government has access to a constant barrage of data from sensors, satellites, social media, mobile communications, email, RFID, and enterprise applications. As a result, leaders are faced with capturing, ingesting, analyzing, storing, distributing, securing the data, and transforming it into meaningful, valuable information.

Since 2000, the amount of information the federal government captures has increased exponentially. In 2009, the U.S. Government produced 848 petabytes of data[1] and U.S. healthcare data alone reached 150 exabytes[2] . Five exabytes (10^18 gigabytes) of data would contain all words ever spoken by human beings on earth. At this rate, Big Data for U.S. healthcare will soon reach zetabyte (10^21 gigabytes) scale and soon yottabytes (10^24 gigabytes).

Yet, the mind-boggling volume of data that the federal government receives makes information overload a fundamental challenge. In this expansion of data, there exists new information that either has not been discoverable, or simply did not exist before. The question is how to effectively capture new insight. Big Data properly managed, modeled, shared, and transformed provides an opportunity to extract new insights, and make decisions in a way simply not possible before now. Big Data provides the opportunity to transform the business of government by providing greater insight at the point of impact and ultimately better serving the citizenry, society and the world.

1 Source: IDC, US Bureau of Labor Statistics, McKinsey Global Institute Analysis

2 Roger Foster, "How to Harness Big Data for Improving Public Health," Government Health IT, April 3, 2012, at http://www.govhealthit.com/news/how-harness-big-data-improving-public-health

Simply put, government leaders find themselves stuck between a rock and a hard place while facing ever-intensifying mission and business challenges, the explosion in the data available, and outmoded, out dated information management capabilities that simply limit their ability to respond. Some of the questions government leaders face include:

- How do I capture, manage and exploit all this new data?
- How do I secure and govern it?
- How do I improve cross-organizational information sharing for broader connected intelligence?
- How do I build trust in the data, through greater understanding of provenance and lineage tied back to validated, trusted sources?
- What advanced visualization techniques, tools, and formats are available for presenting information to enable quick analysis and to create new insights?
- Finally, how do I bridge the gap in talent and human capital to take advantage?

Big Data Definition

Although the term "Big Data" has become increasingly common, its meaning is not always clear. For the purposes of this report, the Commission tapped its collective experience, interviewed government leaders from across the ecosystem, and arrived at the definition below:

> "Big Data is a term that describes large volumes of high velocity, complex and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management, and analysis of the information."

Characteristics of Big Data

Big Data is often characterized by three factors: volume, velocity, and variety. Fifteen percent of the information today is structured information, or information that is easily stored in relational databases of spreadsheets, with their ordinary columns and rows. Unstructured information, such as email, video, blogs, call center conversations, and social media, makes up about 85% of data generated today and presents challenges in deriving meaning with conventional business intelligence tools. Information-producing devices, such as sensors, tablets, and mobile phones continue to multiply. Social networking is also growing at an accelerated pace as the world becomes more connected. Such information sharing options represents a fundamental shift in the way people, government and businesses interact with each other.

The characteristics of Big Data will shape the way government organizations ingest, analyze, manage, store, and distribute data across the enterprise and across the ecosystem. Table 1 illustrates characteristics of Big Data that more completely describe the difference of "Big Data" from the historical perspective of "normal" data.

Table 1: Characteristics of Big Data

Characteristic	Description	Attribute	Driver
Volume	The sheer amount of data generated or data intensity that must be ingested, analyzed, and managed to make decisions based on complete data analysis	According to IDC's Digital Universe Study, the world's "digital universe" is in the process of generating 1.8 Zettabytes of information - with continuing exponential growth – projecting to 35 Zettabytes in 2020	Increase in data sources, higher resolution sensors
Velocity	How fast data is being produced and changed and the speed with which data must be received, understood, and processed	• Accessibility: Information when, where, and how the user wants it, at the point of impact • Applicable: Relevant, valuable information for an enterprise at a torrential pace becomes a real-time phenomenon • Time value: real-time analysis yields improved data-driven decisions	• Increase in data sources • Improved thru-put connectivity • Enhanced computing power of data generating devices
Variety	The rise of information coming from new sources both inside and outside the walls of the enterprise or organization creates integration, management, governance, and architectural pressures on IT	• Structured – 15% of data today is structured, row, columns • Unstructured – 85% is unstructured or human generated information • Semistructured – The combination of structured and unstructured data is becoming paramount. • Complexity – where data sources are moving and residing	• Mobile • Social Media • Videos • Chat • Genomics • Sensors
Veracity	The quality and provenance of received data	The quality of Big Data may be good, bad, or undefined due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations	Data-based decisions require traceability and justification

Mission / Value of Big Data

The manner in which Big Data can be used to create value across the government and in the global economy is broad and far reaching. We are at the cusp of a tremendous wave of innovation, productivity, and growth – all driven by Big Data as citizens, companies, and government exploit its potential.

But, why should this be the case now? Hasn't data always been a part of the impact of information technology? Yes, but the scale and scope of value that data can bring is coming to an inflection point, set to expand greatly as the availability of Big Data converges with the ability to harness it and the affordability of doing so. The federal government has a significant opportunity to boost the efficiency and value of investments for citizens at a time when public finances are constrained and likely to remain so, as U.S. and world population ages and diversifies, the economy continues to globalize, and the expectations on the part of the citizenry intensify.

Many key tenets for "Good Government" and Big Data overlap. At its core, Big Data enables government organizations to be smarter to improve the productivity of the enterprise, and to serve the needs of their stakeholders by improving decision-making in individual agencies and across the government ecosystem. The ability to drive critical, accurate insights with speed across a variety of data sets and across a variety of channels (e.g., web portals, internal systems, smart phones, tablets) will dramatically improve a broad range of critical government policies and practices.

Beyond the mission and business value derived directly by the government's capture, management, analysis, and storage of Big Data, such efforts will also create new markets and business opportunities for the private sector. Harnessing Big Data will enable businesses to improve market intelligence, thereby enhancing the value they provide to consumers. It will also help to reduce business uncertainty for small businesses. A recent report by McKinsey Global Institute (MGI) finds that leveraging Big Data for insights can create significant value for the economy, enhancing productivity and competitiveness for companies and the public sector, and creating a substantial economic surplus for consumers.

As Big Data becomes an increasingly valuable asset to the government, the government's embrace of the principles of Big Data will lay the foundation for:

- Replacing or supporting human decision-making with automated algorithms
- Reducing inefficiencies within an agency
- Creating transparency
- Improving performance by enabling experimentation to discover needs and expose variability
- Improving ROI for IT investments
- Improved decision-making and operational intelligence
- Providing predictive capabilities to improve mission outcomes
- Reducing security threats and crime
- Eliminating waste, fraud, and abuse
- Innovating new business models and stakeholder services

The Use of Big Data: Potential Business & Mission Value

Although the Big Data challenge is daunting, it is not insurmountable, and the opportunity is compelling. There are many possibilities and approaches to managing and leveraging Big Data to address the mission of government inclusive of stakeholder requirements. Ultimately, this report will provide guidance for a framework to extract the Big Data needed to analyze and use for effective decision making. This will be the baseline for a continuous feedback process to improve upon the outcomes identified or potentially eliminate programs that are not delivering on the desired outcomes.

The potential applications of Big Data described below serve to illustrate the "Art of the Possible" in the potential value that can be derived. These applications are consistent with the recently published Digital Government Strategy.[3] They require a customer focus and the ability to reuse and leverage data in innovative ways.

Healthcare Quality and Efficiency

The ability to continuously improve quality and efficiency in the delivery of healthcare while reducing costs remains an elusive goal for care providers and payers, but also represents a significant opportunity to improve the lives of everyday Americans. As of 2010, national health expenditures represent 17.9% of gross domestic product, up from 13.8% in 2000.[4] Coupled with this rise in expenditures, certain chronic diseases, such as diabetes, are increasing in prevalence and consuming a greater percentage of healthcare resources. The management of these diseases and other health-related services profoundly affects our nation's well-being.

Big Data can help. The increased use of electronic health records (EHRs) coupled with new analytics tools presents an opportunity to mine information for the most effective outcomes across large populations. Using carefully de-identified information, researchers can look for statistically valid trends and provide assessments based upon true quality of care.

Healthcare Early Detection

Big Data in health care may involve using sensors in the hospital or home to provide continuous monitoring of key biochemical markers, performing real time analysis on the data as it streams from individual high-risk patients to a HIPAA-compliant analysis system. The analysis system can alert specific individuals and their chosen health care provider if the analysis detects a health anomaly, requiring a visit to their provider or a "911" event about to happen. This has the potential to extend and improve the quality of millions of citizens' lives.

3 http://www.whitehouse.gov/sites/default/files/omb/egov/digital-government/digital-government-strategy.pdf

4 http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/Downloads/tables.pdf

Transportation

Through improved information and autonomous features, Big Data has the potential to transform transportation in many ways. The nemesis of many American drivers, traffic jams waste energy, contribute to global warming and cost individuals time and money. Distributed sensors on handheld devices, on vehicles, and on roads can provide real-time traffic information that is analyzed and shared. This information, coupled with more autonomous features in cars can allow drivers to operate more safely and with less disruption to traffic flow. This new type of traffic ecosystem, with increasingly connected "intelligent cars," has the potential to transform how we use our roadways.[5]

Education

Big Data can have a profound impact on American education and our competitiveness in the global economy. For example, through in-depth tracking and analysis of on-line student learning activities – with fine grained analysis down to the level of mouse clicks – researchers can ascertain how students learn and the approaches that can be used to improve learning. This analysis can be done across thousands of students rather than through small isolated studies.[6] Courses and teaching approaches, online and traditional, can be modified to reflect the information gleaned from the large scale analysis.

Fraud Detection – Healthcare Benefits Services

Big Data can transform improper payment detection and fundamentally change the risk and return perceptions of individuals that currently submit improper, erroneous or fraudulent claims. For example, a significant challenge confronting the Centers for Medicare and Medicaid Services (CMS) is managing improper payments under the Medicare Fee-For-Service Program (FFS). The FFS distributes billions of dollars in estimated improper payments.[7] Currently, contractors and employees identify improper payments by selecting a small sample of claims, requesting medical documentation from the provider who submitted the claims, and manually reviewing the claims against the medical documentation to verify the providers' compliance with Medicare's policies.

This challenge is an opportunity to explore a use case for applying Big Data technologies and techniques, to perform unstructured data analytics on medical documents to improve efficiency in mitigating improper payments. Automating the improper payment process and utilizing Big Data tools, techniques and governance processes would result in greater improper payment prevention or recovery. Data management and distribution could be achieved through an image classification workflow solution to classify and route documents. Data analytics and data intelligence would be based on unstructured document analysis techniques and pattern matching expertise.

The benefit is that the culture of submitting improper payments will be changed. Big Data tools, techniques and governance processes would increase the prevention and recovery dollar value by evaluating the entire data set and dramatically increasing the speed of identification and detection of compliance patterns.

5 http://www.forbes.com/sites/toddwoody/2012/09/19/automakers-on-the-road-to-self-driving-cars/
6 http://www.nytimes.com/2012/03/29/technology/new-us-research-will-aim-at-flood-of-digital-data.html
7 http://www.mcknights.com/gao-medicare-spent-48-billion-on-improper-fee-for-service-payments/article/208592/

Cyber Security

Government agencies face numerous challenges associated with protecting themselves against cyber attacks, such as managing the exponential growth in network-produced datat, database performance issues due to lack of ability to scale to capture this data, and the complexity in developing and applying analytics for fraud to cyber data. Agencies continue to look at delivering innovative cyber analytics and data intensive computing solutions. Cyber intelligence and other machine generated data are growing beyond the limits of traditional database and appliance vendors. Therefore, requirements exist for fast data ingestion, data sharing, and collaboration.

Agencies are looking to incorporate multiple streams of data to benefit both human and automated analysis. Cyber data such as host, network, and information from the World Wide Web, are being fused with human oriented information such as psychosocial, political, and cultural information to form a more complete understanding of our adversaries, motives, and social networks. In addition, deep forensics, critical infrastructure protections, Supervisory Control and Data Acquisition (SCADA) security, and insider threat protection are areas of focus for Big Data cyber security capabilities.

Fraud Detection – Tax Collection

By increasing the ability to quickly spot anomalies, government collection agencies can lower the "tax gap" – the difference between what taxpayers owe and what they pay voluntarily – and profoundly change the culture of those that would consider attempting improper tax filings. Most agencies practice a "pay and chase" model, in which they accept returns and often pay out tax refunds, and only ex post facto review a sampling of returns in order to reveal unintentional or intentional underpayment. Big Data offers the ability to improve fraud detection and uncover noncompliance at the time tax returns are initially filed, reducing the issuance of questionable refunds.

Weather

The ability to better understand changes in the frequency, intensity, and location of weather and climate can benefit millions of citizens and thousands of businesses that rely upon weather, including farmers, tourism, transportation, and insurance companies. Weather and climate-related natural disasters result in tens of billions of dollars in losses every year and affect the lives of millions of citizens. Much progress has been made in understanding and predicting weather, but it's far from perfect. New sensors and analysis techniques hold the promise of developing better long term climate models and nearer term weather forecasts.

New Ways of Combining Information – Helping the Unemployed Find Work

Consistent with the Government Digital Strategy of making information available and having a customer focus, Big Data provides an opportunity to develop thousands of new innovative ways of developing solutions for citizens. For example, Big Data may be useful in proactively helping unemployed job seekers find new opportunities by combining their job qualifications with an analysis and mining of available job opportunities that are posted on the Internet (e.g., company Websites, commercial postings). It'll take some innovation, piloting and experimentation to make these new ideas work, but the payoff can be significant for individuals and society.

Conclusions

Government agencies should think about Big Data not as an IT solution to solve reporting and analytical information challenges but rather as a strategic asset that can be used to achieve better mission outcomes, and conceptualized in the strategic planning, enterprise architecture, and human capital of the agency. Through this lens, government agencies should create an ownership structure for the data, treating it like any other asset – one that is valued and secured. Ultimately, agencies should strive to address the following two questions – "How will the business of government change to leverage Big Data?" and "How will legacy business models and systems be disrupted?"

Big Data Case Studies

The Commission has compiled a set of 10 case studies detailing the business or mission challenge faced, the initial Big Data use case, early steps the agency took to address the challenge and support the use case, and the business results. Although the full text of these case studies will be posted at the TechAmerica Foundation Website, some are summarized below.

Table 2 – Case Studies High Level Summary

Agency/Organization/ Company Big Data Project Name	Underpinning Technologies	Big Data Metrics	Initial Big Data Entry Point	Public/User Benefits
Case Studies and Use Cases				
National Archive and Records Administration (NARA) Electronics Records Archive	Metadata, Submission, Access, Repository, Search and Taxonomy applications for storage and archival systems	Petabytes, Terabytes/sec, Semi-structured	Warehouse Optimization, Distributed Info Mgt	Provides Electronic Records Archive and Online Public Access systems for US records and documentary heritage
TerraEchos Perimeter Intrusion Detection	Streams analytic software, predictive analytics	Terabytes/sec	Streaming and Data Analytics	Helps organizations protect and monitor critical infrastructure and secure borders
Royal Institute of Technology of Sweden (KTH) Traffic Pattern Analysis	Streams analytic software, predictive analytics	Gigabits/sec	Streaming and Data Analytics	Improve traffic in metropolitan areas by decreasing congestion and reducing traffic accident injury rates
Vestas Wind Energy Wind Turbine Placement & Maintenance	Apache Hadoop	Petabytes	Streaming and Data Analytics	Pinpointing the optimal location for wind turbines to maximize power generation and reduce energy cost
University of Ontario (UOIT) Medical Monitoring	Streams analytic software, predictive analytics, supporting Relational Database	Petabytes	Streaming and Data Analytics	Detecting infections in premature infants up to 24 hours before they exhibit symptoms
National Aeronautics and Space Administration (NASA) Human Space Flight Imagery	Metadata, Archival, Search and Taxonomy applications for tape library systems, GOTS	Petabytes, Terabytes/sec, Semi-structured	Warehouse Optimization	Provide industry and the public with some of the most iconic and historic human spaceflight imagery for scientific discovery, education and entertainment
AM Biotechnologies (AM Biotech) DNA Sequence Analysys for Creating Aptamers	Cloud-based HPC genomic applications and transportable data files	Gigabytes, 10^7 DNA sequences compared	Streaming Data & Analytics, Warehouse Optimization, Distributed Info Mgt	Creation of a unique aptamer compounds to develop improved therapeutics for many medical conditions and diseases
National Oceanic and Atmospheric Administration (NOAA) National Weather Service	HPC modeling, data from satellites, ships, aircraft and deployed sensors	Petabytes, Terabytes/sec, Semi-structured, ExaFLOPS, PetaFLOPS	Streaming Data & Analytics, Warehouse Optimization, Distributed Info Mgt	Provide weather, water, and climate data, forecasts and warnings for the protection of life and property and enhancement of the national economy
Internal Revenue Service (IRS) Compliance Data Warehouse	Columnar database architecture, multiple analytics applications, descriptive, exploratory, and predictive analysis	Petabytes	Streaming Data & Analytics, Warehouse Optimization, Distributed Info Mgt	Provide America's taxpayers top quality service by helping them to understand and meet their tax responsibilities and enforce the law with integrity and fairness to all
Centers for Medicare & Medicaid Services (CMS) Medical Records Analytics	Columnar and NoSQL databases, Hadoop being looked at, EHR on the front end, with legacy structured database systems (including DB2 and COBOL)	Petabytes, Terabytes/day	Streaming Data & Analytics, Warehouse Optimization, Distributed Info Mgt	Protect the health of all Americans and ensure compliant processing of insurance claims

The case studies represent systems that have been in production. Given their maturity, some case studies identify existing and new challenges that will require even greater evolution of their IT infrastructures and technology to handle the level of compliance, retention and accuracy required by current and new policies and by public and economic needs.

These new challenges create new use cases for needed Big Data solutions. These new use cases demonstrate how Big Data technologies and analytics are enabling new innovations in scientific, health, environmental, financial, energy, business and operational sectors. That said, the men and women involved with the systems described in these case studies and use cases are persevering. They are committed to the positive outcomes that Big Data solutions offer: better protection, increased collaboration, deeper insights, and new prosperity. The use cases will demonstrate that using the current and emerging technologies for Big Data (including cloud-enabled Big Data applications) will drive new solutions to deliver insights and information that benefits both the government and the public, thus enabling real-time decision making.

National Archive and Records Administration (NARA)

The National Archive and Records Administration (NARA) has been charged with providing the Electronic Records Archive (ERA) and Online Public Access systems for U.S. records and documentary heritage. As of January 2012, NARA is managing manage about 142 terabytes (TB) of information (124 TB of which is managed by ERA), representing over 7 billion objects, incorporating records from across the federal agency ecosystem, Congress and several presidential libraries, reaching back to the George W. Bush administration. It sustains over 350 million annual online visits for information. These numbers are expected to dramatically increase as agencies are mandated to use NARA in FY2012.

In addition to ERA, NARA is currently struggling to digitize over 4 million cubic feet of traditional archival holdings, including about 400 million pages of classified information scheduled for declassification, pending review with the intelligence community. Of that backlog, 62% of the physical records stored run the risk of never being preserved.

The NARA challenge represents the very essence of Big Data – how does the agency digitize this huge volume of unstructured data, provide straightforward and rapid access, and still effectively governing the data while managing access in both classified and declassified environments.

NARA has adopted an approach that put it on the path to developing the Big Data capability required to address its challenge. This approach combines traditional data capture, digitizing, and storage capabilities with advanced Big Data capabilities for search, retrieval, and presentation, all while supporting strict security guidelines. Dyug Le, Director of ERA Systems Engineering writes, “It is best that the Electronic Records Archive be built in such a way so as to fit in a technology ecosystem that can evolve naturally, and can be driven by the end users in ways that naturally ride the technology waves.” [8] The result is faster ingestion and categorization of documents, an improved end user experience and dramatically reduced storage costs. NARA notes the importance of record keeping in their drive toward electronic adoption and the cost benefit. It states, “Electronic records can be duplicated and protected at less cost than paper records.” [9]

8 http://ddp.nist.gov/workshop/ppts/01_05_Dyung_Le%20US_DPIF_NIST%20Digital%20Preservation%20Workshop.pdf
9 http://www.archives.gov/records-mgmt/policy/prod1afn.html

Royal Institute of Technology of Sweden (KTH)

Researchers at KTH, Sweden's leading technical university, wanted to gather in real-time a wide array of data that might affect traffic patterns, in order to better managed congestion. This real-time sensor data includes GPS from large numbers of vehicles, radar sensors on motorways, congestion charging, weather and visibility etc. The challenge was collecting the wide variety of data at high velocity and assimilating it in real time for analysis.

Collected data is now flowing into a commercial off-the-shelf (COTS) Streams Analytics software, a unique software tool that analyzes large volumes of streaming, real-time data, both structured and unstructured. The data is then used to help intelligently identify current conditions, and estimate how long it would take to travel from point to point in the city, offer advice on various travel alternatives, such as routes, and eventually help improve traffic in a metropolitan area.

The KTH Big Data Deployment:

- Uses diverse data, including GPS locations, weather conditions, speeds and flows from sensors on motorways, incidents and roadworks

- Enters data into the Streams Analytics software, which can handle all types of data, both structured and unstructured

- Handles, in real time, the large traffic and traffic-related data streams to enable researchers to quickly analyze current traffic conditions and develop historical databases for monitoring and more efficient management of the system

The result has been a decrease in traffic congestion and accidents in the target cities. KTH is now looking to expand the capability to support routing of emergency services vehicles.

Vestas Wind Energy

Since 1979, this Danish company has been engaged in the development, manufacture, sale, and maintenance of wind power systems to generate electricity. Today, Vestas installs an average of one wind turbine every three hours, 24 hours a day, and its turbines generate more than 90 million megawatt-hours of energy per year, enough electricity to supply millions of households.

Making wind a reliable source of energy depends greatly on the placement of the wind turbines used to produce electricity in order to optimize the production of power against wear and tear on the turbine itself.

For Vestas the process of establishing a location starts with its wind library, which combines data from global weather systems with data collected from existing turbines. Data is collected from 35,000 meteorological stations scattered around the world and from Vestas's turbines. The data provides a picture of the global flow scenario, which in turn leads to mesoscale models that are used to establish a huge wind library that can pinpoint the weather at a specific location at a specific time of day.

The company's previous wind library provided detailed information in a grid pattern with each grid measuring 27x27 kilometers (about 17x17 miles). Using computational fluid dynamics models, Vestas engineers can then bring the resolution down even further—to about 10x10 meters (32x32 feet)—to establish the exact wind flow pattern at a particular location. However, in any modeling scenario, the more data and the smaller the grid area, the greater the accuracy of the models. As a result, Vestas wanted to expand its wind library more than 10 fold to include a larger range of weather data over a longer period of time.

To succeed, Vestas uses one of the largest supercomputers worldwide, along with a new Big Data modeling solution, to slice weeks from data processing times and support 10 times the amount of data for more accurate turbine placement decisions. Improved precision provides Vestas customers with greater business case certainty, quicker results, and increased predictability and reliability in wind power generation.

- Reduces response time for wind forecasting information by approximately 97 percent (from weeks to hours) to help cut development time

- Improves accuracy of turbine placement with capabilities for analyzing a greater breadth and depth of data

- Lowers the cost to customers per kilowatt hour produced and increases customers' return on investment

- Reduces IT footprint and costs, and decreases energy consumption by 40 percent

University of Ontario Institute of Technology

The rapid advance of medical monitoring technology has done wonders to improve patient outcomes. Today, patients routinely are connected to equipment that continuously monitors vital signs, such as blood pressure, heart rate and temperature. The equipment issues an alert when any vital sign goes out of the normal range, prompting hospital staff to take action immediately.

Many life-threatening conditions do not reach critical level right away, however. Often, signs that something is wrong begin to appear long before the situation becomes serious, and even a skilled and experienced nurse or physician might not be able to spot and interpret these trends in time to avoid serious complications. One example of such a hard-to-detect problem is nosocomial infection, which is contracted at the hospital and is life threatening to fragile patients such as premature infants. According to physicians at the University of Virginia, an examination of retrospective data reveals that, starting 12 to 24 hours before any overt sign of trouble, almost undetectable changes begin to appear in the vital signs of infants who have contracted this infection. Although the information needed to detect the infection is present, the indication is very subtle; rather than being a single warning sign, it is a trend over time that can be difficult to spot, especially in the fast-paced environment of an intensive care unit.

The University of Ontario's Institute of Technology partnered with researchers from a prominent technology firm that was extending a new stream-computing platform to support healthcare analytics. The result was Project Artemis -- a highly flexible platform that aims to help physicians make better, faster decisions regarding patient care for a wide range of conditions. The earliest iteration of the project is focused on early detection of nosocomial infection by watching for reduced heart rate variability along with other indications.

Project Artemis is based on Streams analytic software. An underlying relational database provides the data management required to support future retrospective analyses of the collected data.

The result is an early warning that gives caregivers the ability to proactively deal with potential complications—such as detecting infections in premature infants up to 24 hours before they exhibit symptoms. This system:

- Holds the potential to give clinicians an unprecedented ability to interpret vast amounts of heterogeneous data in real time, enabling them to spot subtle trends

- Combines physician and nurse knowledge and experience with technology capabilities to yield more robust results than can be provided by monitoring devices alone

- Provides a flexible platform that can adapt to a wide variety of medical monitoring needs

NASA Johnson Space Center

As the nucleus of the nation's astronaut corps and home to International Space Station (ISS) mission operations, NASA Johnson Space Center (JSC) plays a pivotal role in surpassing the physical boundaries of Earth and enhancing technological and scientific knowledge to benefit all of humankind. NASA JSC manages one of the largest imagery archives in the world and has provided industry and the public with some of the most iconic and historic human spaceflight imagery for scientific discovery, education and entertainment. If you have seen it at the movies or on TV, JSC touched it first.

NASA's imagery collection of still photography and video spans more for than half a century: from the early Gemini and Apollo missions to the Space Station. This imagery collection currently consists of over 4 million still images, 9.5 million feet of 16mm motion picture film, over 85,000 video tapes and files representing 81,616 hours of video in analog and digital formats. Eight buildings at JSC house these enormous collections and the imagery systems that collect, process, analyze, transcode, distribute and archive these historical artifacts. NASA's imagery collection is growing exponentially, and its sheer volume of unstructured information is the essence of Big Data.

NASA's human spaceflight imagery benefits the public through the numerous organizations that create media content for social and public consumption. It is also used by the scientific and engineering community to avoid costly redesigns and to conduct scientific and engineering analysis of tests and mission activities conducted at NASA JSC and White Sands Test Facility.

NASA has developed best practices through technologies and processes to:

- Comply with NASA records retention schedules and archiving guidance
- Migrate imagery to an appropriate storage medium and format destined for the National Archives
- Develop technology to digital store down-linked images and video directly to tape libraries

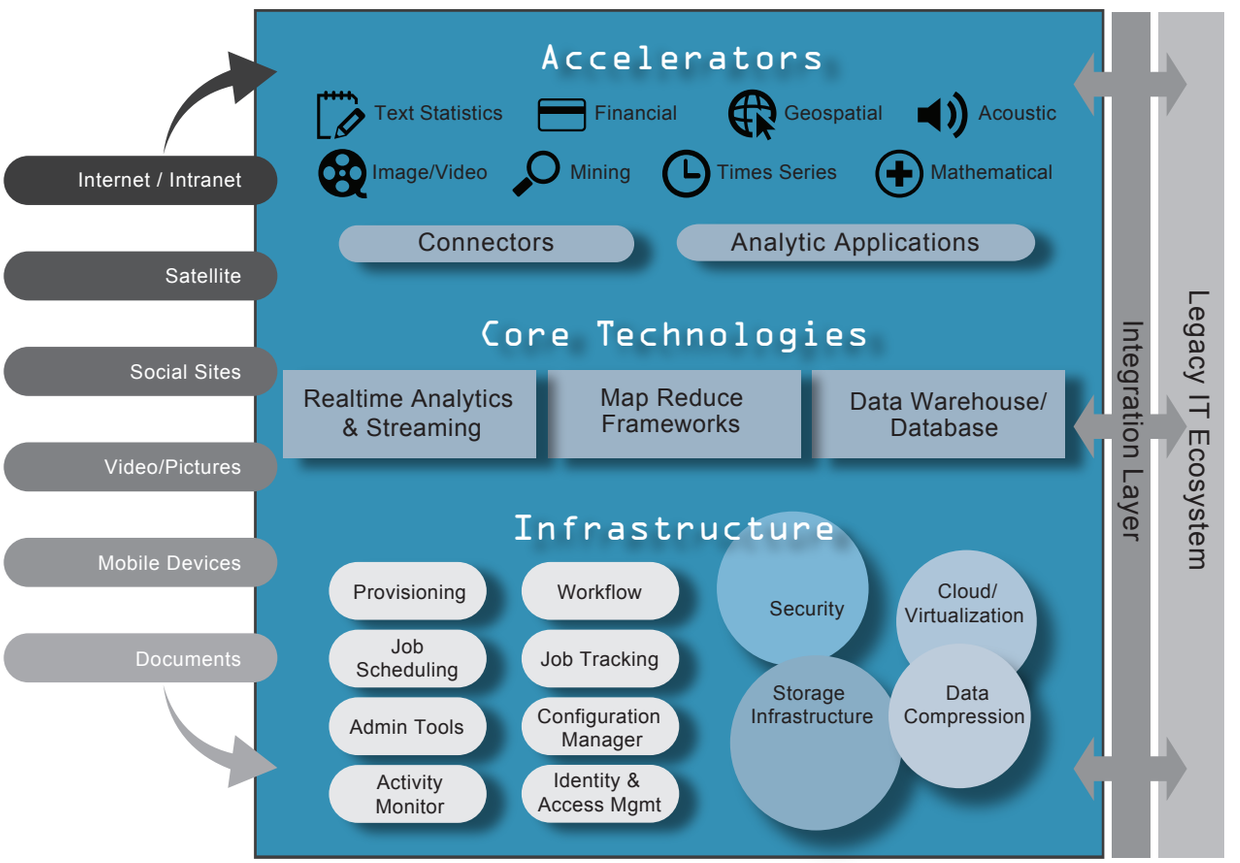
Key lessons learned have revolved around data lifecycle management and NASA's Imagery Online records management tool which contributes to the state of the art in records management.

Technology Underpinnings

Introduction

Powerful advances in new information management and business analytics technologies, like map reduce frameworks (such as Hadoop, and Hadoop-like technologies), stream analytics, and massive parallel processing (MPP) data warehouses, have been proven in deployment to support government agencies in harnessing the value from the increased volume, variety and velocity that characterize Big Data. This section describes the different capabilities provided by these technologies, and how they are combined in different ways to provide unique solutions. No single technology is required for a "Big Data Solution" – they are not "must haves" – as initial Big Data Deployments are unique to the individual agency's business imperatives and use cases. The technologies can be placed, however, into the context of a broader enterprise Big Data model. The model below highlights the ecosystem of technologies that can be used to support Big Data solutions, coupling legacy investments to new technologies. As a result, the technologies listed in the model are not all new, but can be used as part of the solution set (see Figure 1).

Figure 1 Big Data Enterprise Model



Big Data Infrastructure

A suitable technology infrastructure is a key prerequisite for embarking on a successful Big Data strategy. The proper infrastructure is required to leverage the data that originates from the varied applications and data sources. Many government agencies operate a diverse collection of systems based on a wide variety of technology architectures. Adjustments in data center infrastructure may be necessary to implement a Big Data platform. For example, additional dedicated data storage may be necessary to manage massive amounts of unstructured information. Network bandwidth is also a concern of current environments.

Big Data Core Technologies

Real-Time Analytics and Streaming performs the analysis of massive volumes of streaming data, frequently with millisecond response times. The data can originate from many sources: sensors, social media sites, digital pictures, video, transaction records, and communications surveillance and has a wide variation in form. It is particularly ideal when users need real-time decision support for high velocities of data. The characteristics of real-time analytics and streaming capabilities include rule-based algorithms, large I/O capability, and parallel processing. It does not necessarily require a significant amount of storage for data at rest.

Map Reduce Frameworks are often based upon Hadoop and Hadoop-like technologies. They work by providing parallel processing capabilities that move subsets of the data to distributed servers. The primary use is processing massive amounts of data in a scalable manner. Use cases include business analytics, extraction, transformation and loading (ETL), log analysis, and Web search engines. Most Hadoop adopters rely on the native Hadoop Distributed File System (HDFS) configured with direct attached storage (DAS). HDFS is a distributed file system that provides fault tolerant storage capable of managing distributed data across many nodes. The systems commonly work in a "map and reduce" paradigm that allows for a master server to distribute (i.e., map) the problem to worker servers, collect each server's results, and reduce it to a collective result. These technologies are "scale out" and can accommodate large, unbounded data sets without the format requirements of a relational database management system (RDBMS).

Frequently, the software includes redundancy in the file system (e.g., Hadoop File System) or error checking in the data that can facilitate the use of commodity hardware. As a result, it is well-suited to deal with semi-structured and unstructured data, and is particularly ideal when there is a large amount of data maintained to discover new insights over time and when there's not a firm understanding of the value or use of the data a priori. It is good for finding a needle in the haystack among data that may or may not be "core data" for an agency today. Hadoop and Hadoop-like technologies tend to work with a batch paradigm that is good for many workloads, but is frequently not sufficient for streaming analysis or fast interactive queries.

Data Warehousing can be part of an overall integrated Big Data solution because it is ideal for analyzing structured data from various systems to deliver deep operational insight with advanced in-database analytics. Traditional on-line analytic processing (OLAP) data warehouses have scale-up characteristics, require relational, structured data, and have a ceiling on the size of the data that can be processed (e.g., when joining the data with structured query language), but can be a powerful source of analytical capability that can be integrated to provide significant customer value. Frequently the data in an OLAP-based data warehouse has been pre-processed and is known to be of high quality.

Big Data Accelerators

Accelerators are software applications, connectors, or interfaces that provide value-add capabilities such as implementing analytics or enhancing the ability to integrate. Accelerators can decrease deployment time and speed the time to realizing value from the Big Data investment. Accelerators include:

- Text extraction tools or interfaces to common Text Extraction or Natural Language Processing products
- Financial tools and interfaces to common financial products
- Geospatial support and interfaces for ground station and ingest solutions
- Geospatial integration and dissemination
- Acoustic interfacing support
- Imagery and video mining, marking, monitoring, and alerting capability or interfacing support

Integration Layer

Integration between new Big Data capabilities and legacy IT investments is often necessary for deploying Big Data solutions. Information integration and governance allows agencies to understand, cleanse, transform, manage, and deliver trusted information to the critical business initiatives.

We recommend that agencies develop Big Data governance plans. The plans should be a holistic approach to help guide the agency from an information management perspective and follow these key items:

- Identify data and content that are vital to its mission
- Identify how, when, where, and to whom information should be made available
- Determine appropriate data management, governance, and security practices
- Identify and prioritize the information projects that deliver the most value

A Big Data information plan ensures that the major components of the governance plan are working in unison. There are four components to an effective information plan:

- Information strategy is the vision that guides decisions and helps the organization determine how best to support business goals
- Information infrastructure is the technology and capabilities that are needed to establish a common information framework
- Information governance is the policies and practices that facilitate the management, usage, improvement, and protection of information across the organization
- Road map is a phased execution plan for transforming the organization

In practice, Big Data technologies can be integrated to provide a comprehensive solution for government IT leadership. Big Data solutions are often used to feed traditional data warehousing and business intelligence systems. For example, Hadoop within a Big Data model can be used as a repository for structured, semi-structured (e.g., log file data), and unstructured data (e.g., emails, documents) that feeds an OLAP data warehouse. The analytics and visualization tools pull data from the OLAP data warehouse and render actionable information through reports. However, the fundamental step of cleansing the data prior to loading the OLAP data warehouse is absolutely critical in developing "trusted information" to be used within the analytics.

Technologies are evolving to provide government IT leadership choices of characteristics and costs. For example, NoSQL and Hadoop technologies include the ability to scale horizontally without a pre-defined boundary. These technologies may run on commodity hardware or can be optimized with high-end hardware technology tuned specifically to support Big Data. Similarly, NoSQL and Hadoop have different characteristics and capabilities than traditional RDBMSs and analysis tools. The ideal enterprise data warehouse has been envisaged as a centralized repository for 25 years, but the time has come for a new type of warehouse to handle Big Data. MapReduce, Hadoop, in-memory databases and column stores don't make an enterprise data warehouse obsolete. The new enterprise data warehouse will leverage all of these software technologies in the RDBMS or via managed services. This "logical data warehouse" requires realignment of practices and a hybrid architecture of repositories and services. Software alone is insufficient — it will demand the rethinking of deployment infrastructures as well.

Scalable analytics using software frameworks can be combined with storage designs that support massive growth for cost-effectiveness and reliability. Storage platform support for Big Data can include multi-petabyte capacity supporting potentially billions of objects, high-speed file sharing across heterogeneous operating systems, application performance awareness and agile provisioning. Agencies must consider data protection and availability requirements for their Big Data. In many cases data volumes and sizes may be too large to back up through conventional methods. Policies for data management will need to be addressed as well; the nature of many Big Data use cases implies data sharing, data reuse and ongoing analysis. Security, privacy, legal issues such as intellectual property management and liability, and retention of data for historical purposes need to be addressed.

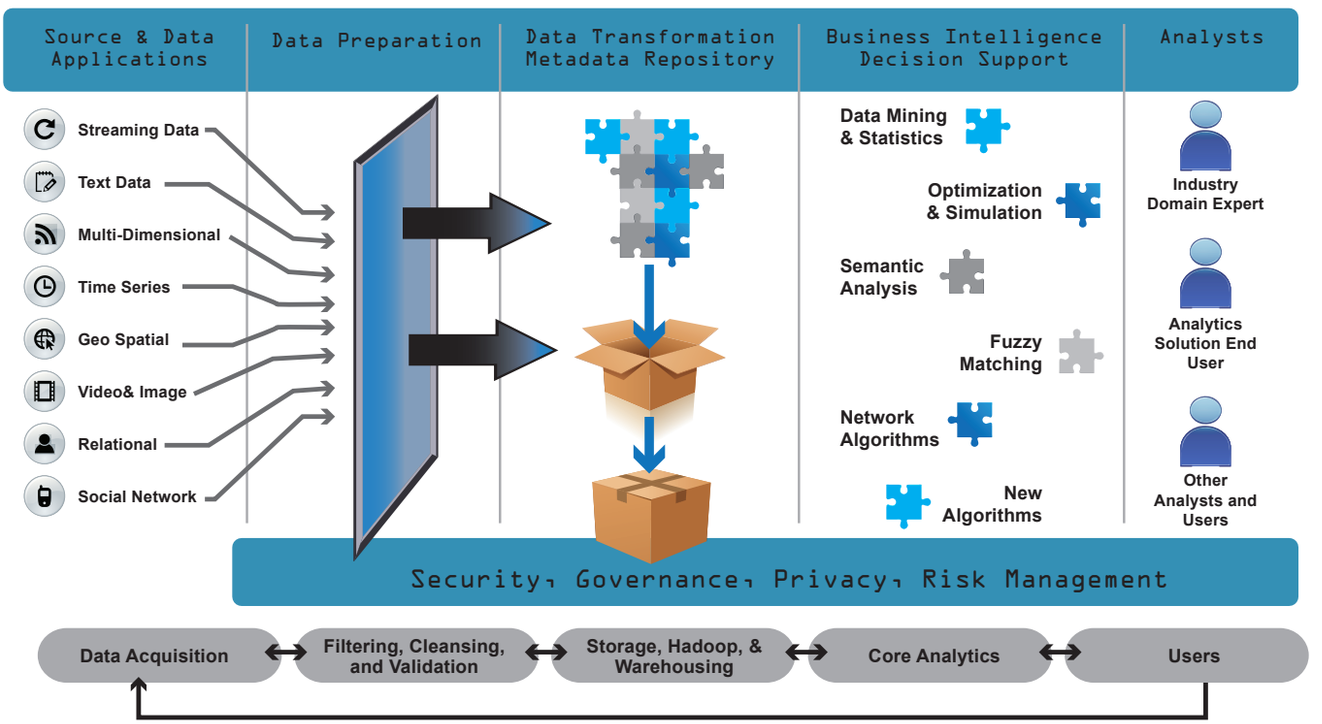
Networking Considerations

Many Big Data implementations require the movement of information across the network and between disparate organizations that operate separate networks. This is especially true when working with near real-time data sources. Current networks may need remediation as information is pulled into storage repositories from multiple sources. New network loads may necessitate a restructuring of agencies' networks to separate traffic used in Big Data solutions from that of normal business operations. The good news is that most government agencies are undergoing significant data center transformations to consolidate operations and standardize on common infrastructures. The impact of standardization, especially at the network and physical level, will allow data center operators to achieve greater flexibility in how they deploy modular data centric solutions. Agencies will gain efficiencies if they ensure data center transformations address these necessary Big Data infrastructure requirements as part of their consolidation strategy.

Cloud Computing Considerations

Consistent with federal CIO policy, Big Data presents federal IT leadership with options for deployment infrastructure that can include cloud computing for development, test, integration, pilots and production. Cloud computing approaches can allow for faster deployment, more effective use of scarce IT resources, and the ability to innovate more quickly. Innovation is enabled through the dynamic use of low cost virtual environments that can be instantiated on demand. This allows organizations to succeed quickly or fail quickly and incorporate their lessons learned. As with any deployment option, cloud computing should be evaluated against the ability to meet application and architecture requirements within the cost structure of the organization.

Figure 2: Notional Information Flow – The Information Supply Chain



A notional flow of information can be a helpful approach in understanding how the technologies underpinning the Big Data Enterprise Model can come together to meet government's needs. This flow follows a simple Understand, Cleanse, Transform and Exploit model. At the end of the day, the key is to map available data sources through the analytic process to the target use cases.

Understand Source Data and Applications

The first step in any information integration and transformation initiative – Big Data or otherwise – is to identify and understand the relevant sources of data, their degree of volume, variety and velocity, and their level of quality. This understanding helps determine the degree of difficulty in accessing the data, the level of transformation required, and the core Big Data technologies that will enable to agency to manage and exploit it.

Data Preparation – Cleansing & Verification

Once an agency understands data sources in the context of the target use case, it can begin to define the method and manner of the data preparation required to support the target use case. For example, unstructured data may require a simple pass through for direct analysis or it may be filtered, cleaned and used for downstream processing. Structured information – such as addresses, phone numbers, and names – may require standardization and verification. Specifics depend on operational and business requirements.

Data Transformation

Once the data fueling a Big Data use case has been cleansed and verified, agencies may consider additional transformation to extract insight and enhance its value. Data may be available for direct analysis (e.g., through Hadoop) or may need additional processing before it can be fully leveraged for the intended use case. Unstructured data, for instance, may be broken down and rendered in a structured format – an agency may want to perform entity extraction to associate organizational or individual names with specific documents. Further, agencies may seek to aggregate data sources, or to understand the non-obvious relationships between them. The goal is trusted information – information that is accurate, complete, and insightful – such that every nugget of information has been extracted from the base data.

Business Intelligence/Decision Support

Once trusted information has been established, agencies can then use the broadest range of analytic tools and techniques to exploit it. These tools and techniques range from the most basic business intelligence capabilities, to more sophisticated predictive analytics, to anomaly detection, content analytics, sentiment analytics, imagery, aural analytics and biometrics. Once the data is brought into the Big Data environment, the critical step is to process it to glean new insights. For example, Hadoop can be used to analyze unstructured data residing on many distributed compute instances or business intelligence tools can be used to analyze a structured data warehouse.

Analysts/Visualization

The final step in the information supply chain is to deliver the new insights created in the previous steps in a manner that most effectively supports the business requirements and use cases. This implies one of two approaches, depending on the users and the business requirement. Either the collection of new insights is provided through a visualization or collaboration tool that enables the users to explore the information, ask questions and uncover new insights; or, the information is simply used to fuel existing work process applications to improve existing processes. Either way, the user needs to be provisioned with data that meets the Big Data needs.

The Path Forward: Getting Started

The TechAmerica Foundation Big Data Commission's report describes the business and mission challenges government agencies face, and the unique role that Big Data can play in addressing these challenges. The report defines the characteristics of Big Data – Volume, Variety, and Velocity – and the key underpinnings required for government leaders to address Big Data challenges and capture the opportunities it offers. This report seeks to illuminate the possibilities by examining potential government use cases, and case studies describing successful deployments. In addition, the Commission has described proven technology and governance principles, methods for getting things done, and ways of addressing privacy and security concerns.

Observations & Lessons Learned

So what have we learned? In our discussion with leaders from across the government ecosystem, and examining the case studies, five central themes have emerged -- each of which coincidentally is well aligned with established engineering best practices:

1. ***Define business requirements:*** Successful Big Data initiatives commonly start with a set of specific and well defined mission requirements, versus a plan to deploy a universal and unproven technical platform to support perceived future requirements. The approach is not "build it and they will come," but "fit for purpose."

2. ***Plan to augment and iterate:*** Successful initiatives favor augmenting current IT investments rather than building entirely new enterprise scale systems. The new integrated capabilities should be focused on initial requirements but be part of a larger architectural vision that can include far wider use cases in subsequent phases of deployment. This approach provides government leaders with lightweight, low trauma solutions to their most immediate pain points within a strategic context that will move them towards an ever greater ability to leverage Big Data.

3. ***Big Data entry point:*** Successful deployments are characterized by three "patterns of deployment" underpinned by the selection of one Big Data "entry point" that corresponds to one of the key characteristics of Big Data. Some initiatives do indeed leverage a combination of these entry points, but experience shows these are the exception.

    • ***Velocity:*** Use cases requiring both a high degree of velocity in data processing and real time decision making, tend to require Streams as an entry point.

    • ***Volume:*** Government leaders struggling with the sheer volume in the data they seek to manage, often select as an entry point a database or warehouse architecture that can scale out without pre-defined bounds.

    • ***Variety:*** Those use cases requiring an ability to explore, understand and analyze a variety of data sources, across a mixture of structured, semi-structured and unstructured formats, horizontally scaled for high performance while maintaining low cost, imply Hadoop or Hadoop-like technologies as the entry point.

4. ***Identify gaps:*** Once an initial set of business requirements have been identified and defined, government IT leaders assess their technical requirements and ensure consistency with their long term architecture. Leaders should identify gaps against the Big Data reference architecture described previously, and then plan the investments to close the gaps.

5. ***Iterate:*** From these Phase I deployments clients typically then expand to adjacent use cases, building out a more robust and unified Big Data platform. This platform begins to provide capabilities that cut across the expanding list of use cases, and provide a set of common services to support an ongoing initiative. Beyond adding in adjacent core capabilities – e.g., Streams, Hadoop, Warehousing and Streams, these services often include:

- Information Integration and Governance
- Privacy & Security
- Common Metadata Management
- Visualization and Discovery

Recommended Road Map for Getting Started

Based on these observations, the Big Data Commission recommends the following five step cyclical approach to successfully take advantage of the Big Data opportunity. These steps are iterative versus serial in nature, with a constant closed feedback loop that informs ongoing efforts. Critical to success is to approach the Big Data initiative from a simple definition of the business and operational imperatives that the government organization seeks to address, and a set of specific business requirements and use cases that each phase of deployment with support. At each phase, the government organization should consistently review progress, the value of the investment, the key lessons learned, and the potential impacts on governance, privacy, and security policies. In this way, the organization can move tactically to address near term business challenges, but operate within the strategic context of building a robust Big Data capability.

Figure 3: Road Map for Getting Started

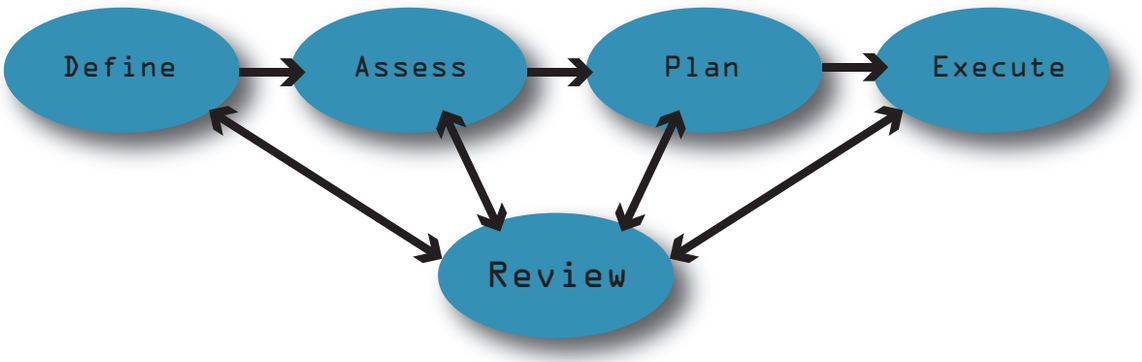


Table 3: Practical Road Map Summary

Define	Define the Big Data opportunity including the key business and mission challenges, the initial use case or set of use cases, and the value Big Data can deliver	• Identify key business challenges, and potential use cases to address • Identify areas of opportunity where access to Big Data can be used to better serve the citizenry, the mission, or reduce costs • Ask – does Big Data hold a unique promise to satisfy the use case(s) • Identify the value of a Big Data investment against more traditional analytic investments, or doing nothing • Create your overall vision, but chunk the work into tactical phases (time to value within 12-18 month timeframe) • Don't attempt to solve all Big Data problems in the initial project – seek to act tactically, but in the strategic context of your key business imperatives
Assess	Assess the organization's currently available data and technical capabilities, against the data and technical capabilities required to satisfy the defined set of business requirements and use cases	• Assess the use case across velocity, variety and volume requirements, and determine if they rise to the level of a Big Data initiative, versus a more traditional approach • Assess the data and data sources required to satisfy the defined use case, versus current availability • Assess the technical requirements to support accessing, governing, managing and analyzing the data, against current capability • Leverage the reference architecture defined in the report above to identify key gaps • Develop an ROI assessment for the current phase of deployment (ROI used conceptually, as the ROI may be better services for customers/citizens and not necessarily a financial ROI)
Plan	Select the most appropriate deployment pattern and entry point, design the "to be" technical architecture, and identify potential policy, privacy and security considerations	• Identify the "entry point" capability as described in the section above • Identify successful outcomes (success criteria) • Develop architectural roadmap in support of the selected use case or use cases • Identify any policy, privacy and security considerations • Plan iterative phases of deployment • Develop program management and acquisitions planning • Identify required skills, resources and staffing • Plan development, test and deployment platforms (e.g., Cloud, HW) • If appropriate, Pilot to mitigate business and technical risk
Execute	The gov't agency deploys the current phase Big Data project, maintaining the flexibility to leverage its investment to accommodate subsequent business requirements and use cases	• Deploy the current phase project plan • Build out the Big Data platform as the plan requires, with an eye toward flexibility and expansion • Deploy technologies with both the flexibility and performance to scale to support subsequent use cases and corresponding data volume, velocity and variety
Review		
The gov't agency continually reviews progress, adjusts the deployment plan as required, and tests business process, policy, governance, privacy and security considerations		• This is a continual process that cuts across the remainder of the roadmap steps • Throughout the assess and planning stages, continually review plans against set governance, privacy, security policies • Assess big data objectives against current Federal, state and local policy • At each stage, assess ROI, and derive lessons learned • Review deployed architecture and technologies against the needs of the broader organization – both to close any gaps, as well as to identify adjacent business areas that might leverage the developing Big Data capability • Move toward Big Data Transformation in a continual iterative process

Public Policy

Accelerating Uptake of Big Data in the Federal Government

As the federal government moves to leverage Big Data, it must look closely at current policies and determine whether they are sufficient to ensure it can maximize its promise. Issues as varied as procurement processes for acquiring Big Data solutions, research and development funding and strategies, and workforce development policies will have a dramatic effect on the success or limitations of federal Big Data efforts. Furthermore, understanding and addressing citizens' expectations on privacy and security is critical for government to implement Big Data solutions successfully. The government should evaluate Big Data policy issues with a view toward removing the unnecessary obstacles to Big Data use, and driving specific actions to accelerate the use and dissemination of the government's data assets. In this section of the report, we address specific strategic policy areas where careful consideration by policy makers can greatly accelerate uptake of Big Data efforts.

The culture of information sharing and decision making needs to grow to include Big Data analysis. Without providing decision-makers the policy; and the incentives to use Big Data for insights and predictions; and the guidance on the use and sharing of information, government will not realize the tremendous value Big Data can offer.

The best private companies today are relentlessly data-driven. They have led the way in industry in measuring their performance carefully and in using hard quantitative information to guide plans and decisions. Unfortunately, in the federal government, daily practice frequently undermines official policies that encourage sharing of information both within and among agencies and with citizens. Furthermore, decision-making by leaders in Congress and the Administration often is accomplished without the benefit of key information and without using the power of Big Data to model possible futures, make predictions, and fundamentally connect the ever increasing myriad of dots and data available. As recognized in a recent Government Accountability Office (GAO) report,[10] Congress may miss opportunities to use performance information produced by agencies. Such data could be leveraged to enact targeted law and policy. So too, the Administration may miss the chance to leverage real-time data as it fulfills program missions. Both branches of government stand to benefit from enhancing policies directed at the use of Big Data approaches as part of their daily routine.

10 Managing for Results: A Guide for Using the GPRA Modernization Act to Help Inform Congressional Decision Making (gao.gov/assets/600/591649.pdf).

The federal government should examine existing organizational and technical structures to find and remove barriers to greater Big Data uptake and, where needed, take action to accelerate its use. Specifically, the government should:

- Assess if annual reviews should incorporate performance metrics based on whether the use of value-added data collaboration activities would improve department and agency performance. Such metrics could promote data sharing and uptake of Big Data in every-day activities. The Administration has made progress in utilizing a host of data-driven reviews to enhance performance. Building upon the success of the TechStat IT program review process, many agencies have instituted data-driven review programs aimed at overarching business and service delivery challenges. HUDStat, FEMAStat, and similar efforts at the Social Security Administration serve as good examples upon which to build. Agencies use these approaches primarily on a quarterly basis. Their use should be expanded.

- Expand upon requirements in the Digital Government Strategy to share and expose data assets of federal agencies to the public and private sector. Requirements to share are beginning to take hold across federal agencies, but stronger requirements and evaluation of agency compliance will help drive further efforts ahead.

- Name a single official both across government and within each agency to bring cohesive focus and discipline to leveraging the government's data assets to drive change, enhance performance, and increase competitiveness. To be certain the value of data is properly harnessed across organizational boundaries, Big Data should not be seen as the exclusive domain of the IT organization or CIO.

- Examine innovative agency approaches to focus on Big Data organizationally, such as the Federal Communications Commission's (FCC) decision to appoint a Chief Data Officer at the FCC, and whether such approaches would be appropriate at other agencies.[11]

These efforts should take a strategic view of the importance of leveraging Big Data for new ways of optimizing operations. It will also be important to clearly define roles and responsibilities to ensure that the discipline of Big Data is not overlooked, but rather is integrated into daily agency operations and decision-making.

11 www.fcc.gov/data/chief-data-officers

Education and Workforce Development

The ability to glean new insights from the massive amounts of data demands new skills and "out-of-the-box" thought processes. At the same time, competition for these new skills and key talent is increasing. Growing demand and competition for top talent, particularly in science, technology, engineering, math, and analytic fields has made hiring and workforce retention increasingly difficult. Recent workforce analyses indicate that there are looming talent shortfalls in key IT disciplines, including those required to implement and manage evolving technologies, such as those associated with Big Data. We recommend a strategy to leverage new talent, while increasing and retaining existing talent with Big Data skills.

Leveraging New Talent

Our greatest source of new thinkers is the collection of students coming from colleges and universities around the country. For government to leverage this source of creative thinkers, the internship programs that currently exist could be expanded to include areas of specific data analytical focus that demand new ways of addressing difficult problems by instituting project focused internship assignments. The additional advantage of such a program would be the increasing connection of government internal issues with people who, as they go out into private industry, can leverage their learning and experience as interns to identify and bring back to government new innovations and opportunities for applying the data that abounds across government agencies.

Increasing Talent

The White House Office of Management and Budget (OMB) should create a formal career track for IT managers and establish an IT Leadership Academy to provide Big Data and related IT training and certification. This academy should foster intellectual curiosity – a trait required to shift workforce focus from transition processing to analytic thinking and continual improvement. To complement these OMB initiatives, agencies should cross-train IT managers to understand functional disciplines so that Big Data and related IT solutions can be effectively utilized to support the mission. We recommend that the IT Leadership Academy be organized using established and proven models of professional development, competency-based education, and in collaboration with select academic centers. These centers can be selected based on their interest in the Academy and their accomplishments in professional development.

To help acquisition and IT personnel understand, adopt, and use Big Data solutions, government agencies, companies, and academia should develop and disseminate educational resources and programs aimed at educating the federal workforce on the technical, business, and policy issues associated with the acquisition, deployment, and operation of Big Data solutions.

We also recommend a broader coalition between Big Data Commission members, academic centers, and professional societies to articulate and maintain professional and competency standards for the field of Big Data. Such standards will guide colleges and universities across the country to develop relevant programs, thus increasing the pool of qualified students for Big Data related roles.

Research and Development Considerations

The Administration's announcement of $200 million in Big Data R&D funding demonstrates its commitment to the transformative nature of Big Data solutions. These investments are a useful start, but much more is needed. In much the same way that sustained federal investment in the technologies underlying the Internet spawned an entirely new technology and economic ecosystem, continued and aggressive investment in the Big Data discipline is critical to the development of new tools, economic models, and educational approaches to advancing Big Data uptake and utilization.

The Commission recommends that the OSTP further develop a national research and development strategy for Big Data that encourages research into new techniques and tools, and that explores the application of those tools to important problems across varied research domains. Such research could be facilitated, and cost could be optimized, by considering the establishment of experimental laboratories within which agencies could explore and test new Big Data technologies without the need to spend funds on their own. This strategy should not be limited to the major research organizations in the federal government, but rather focus on the roles and skill sets of all levels of government and create incentives for the private sector to innovate and develop transformative solutions.

The R&D strategy should also move beyond technical attributes and solutions, to address the wide range of domains where improvements are needed. Among the issues we recommend including in the national Big Data R&D strategy are the following:

A. Education and training advancements, including novel approaches to curriculum development and data-intensive degree programs and scholarships to expand the preparation of new generations of data scientists;

B. New management and analytic tools to address increasing data velocity and complexity;

C. Novel privacy-enhancing and data management technologies;

D. Development of new economic models to encourage data sharing and monetization of Big Data solutions and collaborative data science efforts;

E. Continued research and development of advanced computing technologies that can effectively process, not only the vast amounts of data being continually generated, but also the various types of data; and

F. Policy guidelines that mandate that agencies identify clear objectives for use of Big Data technologies with a focus on metrics that can reduce the cost of operations.

The Commission has spent time with the cross-government Big Data Steering Group operating under the auspices of the National Coordination Office for Networking and Information Technology Research and Development (NITRD). We applaud their efforts to explore additional avenues to innovation, including competitions and prizes, and encourage the R&D strategy to address this approach explicitly. The Administration has achieved success with internal government competitions, and we encourage efforts of a similar nature.

Privacy Issues

As the earlier best practice section reveals, good data protection practices are in place, as evidenced by the existing policies and industry best practices related to privacy and security by design context, accountability, and transparency. The Commission does not believe that realizing the promise of Big Data requires the sacrifice of personal privacy. Accordingly, education and execution on existing privacy protections and procedures must be part of any Big Data project.

There are over 40 laws[12] that provide various forms of privacy protections. Some, such as the Privacy Act of 1974, provide protections to personal data used by government; others are more sector-specific, such as the Health Insurance Portability and Accountability Act (HIPAA), and the Financial Modernization Act (Gramm-Leach-Bliley Act 1999), which concern health and financial information, respectively. Further guidance from OMB, and greater collaboration with industry and stakeholders, on applying these protections to current technology and cultural realities could help accelerate the uptake of Big Data. Specifically, we recommend that OMB strongly collaborate with industry and advocacy groups and issue additional guidance that addresses:

1. Ways to simplify the understanding of privacy management obligations through the issuance of one guiding set of principles and practices to be applied by all agencies. Along with the streamlining of access to what management practices are required, this guidance also should explicitly note the penalties for violating privacy requirements.

2. The government's use of data that is not part of a Privacy Act System of Records.

3. The growing concern about the use, aggregation, sharing, retention, and deletion of data with an eye toward identifying best practices that may be needed to realize Fair Information Practice Principles (FIPPs).

4. The need to recognize and promote citizen confidence by communicating clearly that, for many of Big Data projects, the data being aggregated are non-regulated, de-identified information with no need to re-identify to derive benefits. Of particular note should be the fact that data sets can be used without personally identifiable information (PII) and still yield positive results (e.g., disease spread predictions).

5. The use of clear public policies regarding notice and access for Big Data to enable choice and transparency, as articulated by the FIPPs. Along these lines, stakeholders could explore a "layered notice" approach, where basic information is available for all, with notice/consent/access requirements scaled to instances where more information is sought.

Most importantly, leadership at all levels of government is needed on privacy as over time with relatively little guidance, agencies have taken action, or not, as they best saw fit. Privacy guidance from OMB should be updated regularly to take into account the latest technologies, cultural attitudes and norms, and government needs. Fundamentally, any policy must be drafted carefully to enable both innovation and appropriate data protection.

12 http://healthit.hhs.gov/portal/server.pt/gateway/PTARGS_0_11113_911059_0_0_18/Federal%20Privacy%20Laws%20Table%202%2026%2010%20Final.pdf

Removing Barriers to Use through Procurement Efficiencies

The acquisition process for IT increasingly involves multiple buying vehicles with low minimum government commitments, compelling industry to participate through numerous, duplicative contracting channels to maximize the ability to win government business. A recent Bloomberg study indicates that the number of multiple award contracts in the federal government has more than doubled from 2007 to 2011.[13] This process continues to inflate administrative costs with no appreciable value added, and threatens to reduce incentives to participate in the government market.

Big Data continues to evolve rapidly, driven by innovation in the underlying technologies, platforms, and analytic capabilities for handling data, as well as changes in user behavior. Given this rapid evolution, coupled with the scarcity of funds competing for public sector initiatives, the proliferation of contract vehicles, especially at the agency level, creates potential problems for its adoption by government. The process- and cost-intensive nature of the current contracting trend creates a barrier to industry vendors offering cutting edge solutions, and at the same time, it hinders federal government efforts to deploy Big Data capabilities in standard approaches across the landscape of disparate federal agencies.

The government should avoid Big Data contract vehicle duplication by promoting an express preference for the use of existing successful government-wide vehicles for Big Data solutions and use of the minimum number of buying vehicles necessary to fulfill its needs, including the Federal Supply Schedules program. Although that program would have to be modified to address some shortcomings, which can restrict its use in the acquisition of complex solutions, it could supplement GWACs, and together, these existing contract vehicles could be leveraged fully before any consideration is given to creating new contracting vehicles. While having the right contracting vehicles is important, the government model for buying technology also needs to be updated to allow for a shift, when appropriate, from a capital expenditure (CapEx) to an operating expense (OpEx) model to pay for usage.

The government has within its management purview the means to craft buying solutions for Big Data without turning to new contracting vehicles that will add cost and administrative burden to the implementation of Big Data solutions. Channeling Big Data solutions through the minimum number of contract vehicles necessary would allow maximum integration, strategic sourcing, governance, and standardization.

13 "The Future of Multiple-Award Contracting" Conference, Presentation by Brian Friel, Bloomberg Government, July 18, 2012.

Conclusion

We live in an exciting time, when the scale and scope of value that data can bring is coming to an inflection point, set to expand greatly as the availability of Big Data converges with the ability to affordably harness it. Hidden in the immense volume, variety and velocity of data that is produced today is new information – facts, relationships, indicators and pointers -- that either could not be practically discovered in the past, or simply did not exist before. This new information, effectively captured, managed, and analyzed, has the power to change every industry including cyber security, healthcare, transportation, education, and the sciences.

To make data a strategic asset that can be used to better achieve mission outcomes, data should be included in the strategic planning, enterprise architecture, and human capital of each agency. These precepts are embodied in Digital Government Strategy, a primary component of which is to "unlock the power of government data to spur innovation across our nation and improve the quality of services for the American people."

Within this report, we have outlined the steps each government agency should take toward adopting Big Data solutions, including the development of data governance and information plans. The Commission recommends that the OSTP further develop a national research and development strategy for Big Data that encourages research into new techniques and tools, and that explores the application of those tools to important problems across varied research domains. We recommend that the OMB strongly collaborate with industry and advocacy groups and issue additional guidance that addresses privacy issues.

Because of the importance of data in the digital economy, the Commission encourages each agency to follow the FCC's decision to name a Chief Data Officer. To generate and promulgate a government-wide data vision, to coordinate activities, and to minimize duplication, we recommend appointing a single official within the OMB to bring cohesive focus and discipline to leveraging the government's data assets to drive change, improve performance, and increase competitiveness.

By following these recommendations, the early successes we described at NARA, NASA, NOAA, IRS, CMS, and the Department of Defense can be expanded and leveraged across government to reduce cost, increase transparency, and enhance the effectiveness of government ultimately better serving the citizenry, society, and the world.

Acknowledgements

TechAmerica Foundation gratefully acknowledges the contributions of the Commissioners, their colleagues and staff to the successful completion of this report. The Commission interviewed and collected input and feedback from a variety of federal government representatives. We express our thanks for their participation in this effort and the valuable knowledge and innumerable years of experience they shared to inform the recommendations of this report.

- **Jeff Butler**, Internal Revenue Service
- **Mark Hogle**, Centers for Medicare and Medicaid Services
- **Suzanne Iacono**, National Science Foundation
- **Tom Kalil**, Office of Science and Technology Policy
- **Mark Loper**, Centers for Medicare and Medicaid Services
- **Vish Sankaran**, Centers for Medicare and Medicaid Services
- **Lisa Schlosser**, Office of Management and Budget
- **George Strawn**, National Coordination Office for Networking and Information Technology Research and Development
- **Larry Sweet**, NASA Johnson Space Center
- **Tony Trenkle**, Centers for Medicare and Medicaid Services
- **Steve Van Roekel**, Office of Management and Budget
- **Wendy Wigen**, National Coordination Office for Networking and Information Technology Research and Development
- **Marc Wynne**, Centers for Medicare and Medicaid Services

Deputy Commissioners

Phil Agee
Attain

Amr Adwallah
Cloudera

Joe Barr
ID Analytics

Greg Bateman
Microsoft

Chirayu Desai
CGI

Roslyn Docktor
IBM

Carolyn Eichler
CSC

John Ellis
Dell

Jackie Ervin
Globanet

Greg Gardner
NetApp

Steven Garrou
Savvis, A CenturyLink Company

Tom Guarini
MicroStrategy

Samuel Henry
MEI Technologies

Darren House
GTSI Corp.

David Jonker
SAP

Soo Kim
TASC

Caron Kogan
Lockheed Martin
IS&GS

Jake Kolojejchick
General Dynamics
C4 Systems

Tim Paydos
IBM

Brendan Peter
CA Technologies

Bethann Pepoli
EMC Corporation

Larry Pizette
Amazon Web Services

Parvathy Ramanathan
Poplicus, Inc.

Brian Reynolds
Grant Thornton LLP

LiMing Shen
Wyle

Seshubabu Simhadri
GCE

Tom Sisti
SAP

Andras Szakal
IBM

Peter Thawley
SAP

John Topp
Splunk

Diana Zavala
Hewlett-Packard

TechAmerica FOUNDATION
601 Pennsylvania Avenue, N.W.
North Building, Suite 600
Washington, D.C. 20004

TechAmerica
FOUNDATION

601 Pennsylvania Avenue, N.W. North Building, Suite 600 Washington, D.C. 20004