CDW·G | PEOPLE WHO GET IT™

PROACTIVE PLANNING FOR BIG DATA

In government, Big Data presents both a challenge and an opportunity that will grow over time.

Executive Summary

Consider this list of government–adopted digital resources: citizen–facing websites, mobile computing, electronic records, cloud services, social networks, smart sensors, IP video, geographic information systems and more. As society has grown increasingly digital, government agencies are harnessing digital technology to operate more efficiently, improve service to citizens and better protect people and places.

With the use of these technologies has come an influx of data. Lots of data. Big Data.

Today, most agencies generate digital information with a clear understanding of how it contributes to their missions — budgets, reports and gigabytes of operational information. But that's just the beginning.

Table of Contents

TWEET THIS!

Spurred by digital government initiatives and advances in technology, federal, state and local agencies now collect terabytes or more of data from new and varied sources: the logs of visitors to government websites; video streams from a growing number of traffic cameras; data from intrusion prevention systems that protect agencies' online assets; and geospatial information on everything from gravesite locations at Arlington National Cemetery to county forest preserves.

Thanks to rapid advances in data storage technology and cloud-based infrastructures, agencies have access to abundant, cost-effective solutions for storing all the data coming in. The challenge comes in managing this data — the Big Data — and using it to make better decisions, discover new ways of doing things, advance science and healthcare, and secure the nation's critical infrastructure.

So what exactly is *Big Data?* In IT circles, it's an integrated set of solutions, including data warehousing, business intelligence and analytics, that allows government staff to connect the dots in the terabytes of data they collect. Getting a handle on Big Data is a challenge, but it's also an opportunity. Answers to pressing constituent needs or long-standing governance problems may be found in those terabytes of data.

What Is Big Data?

New trends in IT are often thought of in terms of leading-edge technology solutions to significant enterprise challenges. In other words, organizations face challenges (for example, improving data center efficiency or providing IT services to remote workers) and IT solutions (such as cloud computing and mobile devices) address them. But one of the biggest trends in IT today, Big Data, is actually named for the challenge it represents, rather than the solution.

At its core, Big Data means lots of data — so much data collected via so many evolving mechanisms that it can be overwhelming. It is increasingly easy for government agencies to store all that data. But Big Data as an IT strategy requires making sense of the data being collected — processing, analyzing and exploiting it for government, partner and constituent gain.

Big Data refers to digital information that is massive and varied, and that arrives in such waves that it requires advanced technology and best practices to sort, process, store and analyze. Organizations that do so effectively can use it to their advantage. Big Data is less about the terabytes than it is about the query tools and business intelligence software needed to make sense of the terabytes.

Three V's

Big Data is often described in terms of three V's: volume, velocity and variety.

Volume: The hard disk drives that stored data in the first personal computers were minuscule compared to today's hard disk drives. Storage volumes have grown because data volumes have grown, whether it's the space required to run a modern operating system or data sets that are used to map genomes. Despite the fact that multiterabyte hard disk drives cost as little as $100, the advent of data such as high-definition video ensures that users will fill up those drives in short order.

Velocity: The challenge isn't just that the world today generates lots of data; it generates data quickly — at high velocity — often in real time. Consider a modern IP video camera used to monitor a bridge crossing. Software can analyze the incoming video data to alert security professionals if a suspicious package appears on the bridge that wasn't in previous video frames.

But to be effective, this analysis must happen as rapidly as the video data enters the system. In this case, velocity pertains not only to how quickly data is generated, but also to how quickly someone interprets and acts upon it.

Another source of high-velocity data is social media. Twitter users are estimated to generate nearly 100,000 tweets every 60 seconds. This comes in addition to almost 700,000 Facebook posts and more than 100 million emails a minute. Somewhere in that deluge is information related to an agency's mission, perhaps from citizens voicing their dissatisfaction or seeking assitance.

Variety: Social media data is also a good example of the variety of information that characterizes Big Data. Social media information, like roughly 80 percent to 90 percent of all data today, is unstructured. It doesn't arrive in neat records that are easily searchable. Unstructured data is often text heavy and doesn't fit neatly into relational tables. Its explosion in recent years has driven the Big Data movement.

Sensors are another massive source of unstructured and semistructured data. Researchers at HP Labs estimate that by 2030, 1 trillion sensors will be in use. These sensors monitor conditions in the physical world, such as weather, energy consumption and environmental surroundings, as well as in cyberspace. Depending on the application, sensors can generate multiple terabytes of data per day.

Structured data is what comes to mind when thinking about traditional databases, filled with customer relationship management (CRM) records, statistics or financial transactions. One of the best opportunities presented by Big Data technologies is to bring together structured and unstructured data to reveal new insights.

Ironically, the challenge of Big Data actually begets even more data. Big Data (especially unstructured data) must be described in a way that software tools, such as business

intelligence, analytics and query tools, can identify and ingest. That's where metadata comes into play.

Metadata is data that describes data, making it discoverable across an enterprise infrastructure or even in the cloud. Agencies must manage metadata as well as the underlying data. The better they manage metadata, the more valuable their Big Data will be.

Government and Big Data

Government's struggle with Big Data is unique. Not only do government agencies deal with all types of data, from scientific research and intelligence to census data and financials (all under the pressure of the three V's), but they also are required to handle data according to government rules and guidance. For example, the E–Government Electronic Records Management Initiative guides agencies on creating electronic versions of all their records (most of which end up unstructured) and transferring them to the U.S. National Archives. Meanwhile, the Federal Information Security Management Act (FISMA) governs how federal agencies protect data. Similarly, several states have laws or regulations governing how they handle public data, such as Massachusetts' Executive Order No. 504.

Moreover, at the same time that digital technology has enabled government to collect massive amounts of data, it's become an important tool for encouraging transparency. As a result, agencies must be able to handle their data efficiently, make sense of what it reveals and share that data openly with the public, in a format that citizens can consume. This can be health statistics, real estate records, meteorological data and more. Many early e–government programs were based on making information available to the public online via web applications.

Finally, along those same lines, government is an important purveyor of data to the research and development, scientific, law enforcement and other communities. If government doesn't embrace Big Data solutions, it could severely hinder scientific progress and government–sponsored security initiatives.

90% The amount of digital data that is unstructured

SOURCE: *The Impact of Big Data on Government*
(*IDC Government Insights,* October 2012)

Big Data, Big Benefits

What potential does Big Data hold for government agencies? By tackling their Big Data challenges and using analytics to unlock key information, agencies can realize a variety of benefits, from improving the way they already utilize IT to discovering new services and capabilities.

Big Data technologies allow groups to play out scenarios under controlled circumstances, customize what–if planning to different organizations, support data–backed decision-making, identify correlations and trends in underlying data and more. By laying the foundation for effective use of Big Data, agencies can:

- **Make better decisions more quickly:** By identifying trends and other insights locked in Big Data, agencies improve their decision–making. By doing it with streaming analytics tools and other technologies to process data generated in real time, the decisions come more quickly. Without these tools and technologies, decision–makers may revert back to merely guessing or decision avoidance altogether.

- **Improve mission outcomes:** Big Data brings with it the ability to predict results and model scenarios based on the data.

- **Identify and reduce inefficiencies:** By working through Big Data, organizations can see where they are taking unnecessary steps, based on the data their processes generate.

- **Eliminate waste, fraud and abuse:** By identifying inefficiencies, organizations can wring out internal waste. Depending on their missions, agencies can also potentially identify and eliminate fraud and abuse by the people or parties they serve or monitor.

- **Improve productivity:** With the right tools, even nontechnical users can work with large data sets to find information, deliver better services or make decisions that support the mission.

- **Boost ROI, cut total cost of ownership (TCO):** At its heart, Big Data solutions are about making better use of the data that IT systems generate, and by extension improving the return on those IT investments. By potentially consolidating data silos and analytics tools, agencies can reduce the TCO for their infrastructures.

- **Enhance transparency and service:** Proper handling and processing of Big Data allows agencies to make data available not only to public– and private–sector partners, but also to the public. This enables citizens to understand what information the government collects. Processing and sharing Big Data also allows agencies to offer information as a service, whether it's online tax records, census information, weather data or more.

- **Reduce security threats and crime:** Analyzing Big Data is key to helping police, homeland security officials, intelligence analysts and others pinpoint patterns and other hidden information to help identify specific threats.

In April 2012, federal agencies were required to detail their strategies for using Big Data. In developing their plans,

agencies made public hundreds of thousands of data sets, covering auto safety, air travel, air quality, workplace safety, drug safety, nutrition, crime, obesity, employment and health-care. They are using this data to improve their operations.

For example, in fiscal 2012, the Department of Health and Human Services made $64.8 billion in improper payments (down from roughly $66 billion in 2011). As part of its Open Government Plan 2.0, the department says it will exploit Big Data to analyze related data sets for information on healthcare expenditures, services and the cost of care in communities around the country.

On the local level, the city of Boston will use data collected from users' smartphones as part of its Street Bump project to create a map of road conditions around the city. This data will provide greater accuracy for repaving initiatives and will save the city more than $100,000 in survey costs.

Barriers to Big Data Success

Like other major IT initiatives, Big Data presents challenges that must be addressed if they're to be successful. This is not to say that agencies need to start these initiatives from scratch. Some challenges may have been overcome already (at least partially) in the course of other IT initiatives.

For example, one significant barrier to implementing an effective Big Data solution is siloed data. More than any other enterprise, government is characterized by departments and agencies that maintain their own independent data stores, which, if integrated properly, can combine to reveal new efficiencies and actionable information.

Data warehousing is one way to help eliminate silos. Efforts over the past decade to better share information among government agencies have already begun to chip away at data silos. But more work must be done, especially as government accumulates new and varied data in new silos.

Data management is another area where agencies have begun to tackle barriers to effective Big Data solutions. For Big Data to be effective, it must be secure and reliable. Therefore, agencies' information assurance (IA) programs are critical. Not only must data be kept safe, but IA solutions must also ensure that the underlying data remains legitimate and unchanged.

Moreover, government agencies must adopt measures to ensure information privacy. For example, Big Data processing can yield insightful results without requiring personally identifiable information. Because of the role of government in people's lives, agencies must be diligent about data privacy or risk a backlash of distrust.

Some other barriers to Big Data success are harder to solve. To start with, agencies must address usability issues by implementing visualization and decision-support tools. It will be hard to realize the benefits of Big Data if agency personnel can't use or make sense of the results.

Big Data from the Top Down

In March 2012, the Obama administration announced $200 million in funding for Big Data research and development. The effort was spearheaded by the White House Office of Science and Technology Policy (OSTP) and several federal agencies, including the Departments of Defense and Energy, the National Institutes of Health, the National Science Foundation and the U.S. Geological Survey.

OSTP Director John Holdren said at the time of the announcement, "In the same way that past federal investments in information technology R&D led to dramatic advances in supercomputing and the creation of the Internet, the initiative we are launching today promises to transform our ability to use Big Data for scientific discovery, environmental and biomedical research, education, and national security."

States are taking advantage of Big Data opportunities too. For example, in May 2012, Massachusetts unveiled an initiative to create a Big Data consortium of leaders from industry and academia and to fund Big Data projects across the state.

Among the federally funded Big Data projects currently under way are:

- **Expeditions in Computing at the University of California, Berkeley:** With a grant from the National Science Foundation, Berkeley researchers are developing the Berkeley Data Analytics Stack, an open-source Big Data solution that seeks to improve on current technologies such as Apache Hadoop and the Hadoop Distributed File System (HDFS), in part by conducting analyses in computer memory.

- **The Defense Advanced Research Projects Agency's XDATA program:** DARPA is developing software tools and techniques to process and analyze Big Data accumulated during Department of Defense missions. Today's military personnel rely heavily on advanced sensors and other systems that generate data.

- **The 1000 Genomes Project:** The National Institutes of Health worked with Amazon Web Services to make hundreds of terabytes of human genomic data available for free to researchers in the cloud, thereby accelerating discovery and genetic research.

Moreover, the data integration and analysis components of Big Data require a fresh set of skills not common in today's IT workers. In a recent survey by the research, analysis and advisory firm IDC, nearly 70 percent of respondents said the skill they most sought to handle their Big Data challenges was experience in advanced or predictive analytics.

82% The percentage of public IT officials who say effective use of Big Data is the wave of the future

SOURCE: *"Big Data and the Public Sector"* (*TechAmerica Foundation,* October 2012)

Getting Ahead of the Deluge

With an understanding of what Big Data means and how it can benefit government, agencies can begin to think differently about this important asset — because Big Data is, indeed, an asset and not just an IT challenge. In its 2012 study, *Demystifying Big Data,* the TechAmerica Foundation's Big Data Commission wrote:

> 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. ... 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?"

In 2010, the Federal Communications Commission appointed chief data officers in each of its offices and bureaus to help the FCC better use and share data, making it the first agency to do so. Others have suggested installing CDOs throughout government. But the title is less important than the job, whether it's the CIO who handles Big Data issues or some other person in a position to enforce Big Data policy and governance.

Because the amount of data collected will continue to grow (and because the potential benefits of analyzing that data are huge), it's important for government agencies to establish a Big Data strategy. But they shouldn't rush into it. There are several steps to be taken first to determine a strategy.

1. Assess the current environment. Organizations need to know what data they already collect, how they collect it and where it resides. They also need to know what condition the data is in so they know how much it must be processed to be useful.

2. Identify business requirements. Without knowing why they want to exploit Big Data, agencies may find it difficult (and wasteful) to pursue a strategy. Studying how other agencies have leveraged Big Data is a good way of understanding the possibilities.

3. Plan to start small. Many agencies already have the building blocks of a strategy in place. The goal should be to grow a Big Data solution out of an existing enterprise architecture, rather than to build a whole new system.

4. Let requirements dictate actions. To start, agencies should map their requirements to the appropriate Big Data solution. If they're dealing with large amounts of data (volume), their first Big Data deployment may be a data warehouse.

If the agency needs to analyze data very quickly (velocity), it may start with real-time analytics, which requires more processing power than storage capacity. If it has identified a need to analyze unstructured data (variety), the agency may want to evaluate solutions such as the open-source Hadoop framework.

5. Identify technical requirements. With an understanding of their existing IT environment and how they'd like to begin using Big Data, government groups can determine where additional infrastructure is needed, if anywhere. This would also be the time to evaluate a cloud-based solution, which would allow them to acquire computing and storage capacity in a pay-as-you-go model without significant upfront investment.

Other Big Data solutions, such as data analysis, are also available in the cloud as a service. At the federal level, the Federal Risk and Authorization Management Program (FedRAMP) authorizes cloud services for use by federal agencies as well as state and local governments.

Agencies must also establish clear data governance policies. In an October 2012 survey of federal and state IT officials by the TechAmerica Foundation, policy and privacy concerns were the most-cited barriers to Big Data adoption. At the federal level, a lack of clear ownership was also seen as an issue. Data governance helps address such concerns and allows agencies to unlock Big Data's potential.

By nature, most Big Data programs comprise disparate sources that need to be integrated. Therefore, the data from those sources must be harmonized, validated and monitored under the same set of rules.

For starters, data governance needs to become an established program, with stakeholders from all groups represented in the Big Data initiative. The program must develop a clear understanding of what data will be included, where it resides, who owns it, who can access it and what they

can do with it. To be effective, a data governance policy should address, among other things:

- Data stewardship
- Data standards
- Architecture and integration
- Risk and risk management
- Data quality
- Metadata
- Privacy, compliance and security
- Information lifecycle management
- Program management

Data governance may evolve holistically as agencies recognize and address issues, such as how to ensure data quality and apply common metadata. But it will be more effective (and will accelerate results) if data governance is part of a formal framework. Data governance takes the guesswork out of identifying and processing useful data and makes it easier to securely introduce new sources as they become available.

---

7 Tips for Tackling Big Data

In 2012, the National Association of State CIOs (NASCIO) published a report on Big Data. Its recommendations apply to government at all levels.

1. Restrict Big Data–driven purchases to those with a strong cost–benefit analysis. Limit them to specific opportunities and not blanket solutions.

2. Consider Big Data investments in light of other necessary initiatives, such as IT modernization.

3. Evaluate investments with an eye toward future sources of data and the need to manage them.

4. Avoid uncoordinated initiatives that may lead to the type of silos that Big Data technologies are meant to solve.

5. Make sure to fit Big Data plans into an enterprise architecture.

6. When evaluating analytics solutions, take into account future data needs, such as more diverse or streaming data.

7. Use stakeholders' interest in Big Data solutions to encourage better data management and data governance practices.

Source: *Is Big Data a Big Deal for State Governments?*
(*NASCIO,* August 2012)

The Big Data Toolbox

Big Data does not necessarily mean a suite of new IT systems that must be procured and spun up. One reason this tech trend has gained momentum in recent years is the maturity of tools that can handle huge data volumes and run on commodity infrastructures.

The foundational elements of a Big Data initiative are the same as they are for many other IT programs: industry-standard servers, networks, storage systems and clustering software. Basically, organizations that have already adopted a "scale out" model for IT infrastructure (increasing computing power by adding hardware resources) already have the experience needed to lay the groundwork for handling Big Data. And organizations that have experience with infrastructure as a service (IaaS) understand another path toward Big Data solutions.

One area of infrastructure that agencies may want to reexamine is storage. Despite the availability of abundant, cost-effective, high-speed data stores, a significant share of government CIOs (37 percent) cite storage as a challenge in dealing with Big Data, according to IDC Government Insights.

With the explosion of data, physical storage is a more strategic asset, in part because the trend line of data creation (upward steeply) is often at odds with the trend line of government IT budgets (downward). Therefore, agencies should right-size their storage systems — ensure they have enough flexibility to scale up their storage to meet Big Data requirements without purchasing capacity they don't need.

In general, organizations embracing Big Data turn to network-attached storage (NAS) solutions, which are easier to scale, install and manage than fixed storage. NAS also offers high utilization rates, meaning agencies are less likely to buy what they don't need.

Broadly speaking, government groups that are considering storage needs to support Big Data initiatives should consider a multitiered approach. This includes tape drives for archiving data, traditional disks for frequently accessed data, solid-state storage for rapid data analysis and cloud storage for very large data sets, or data sets the agency might want to make accessible beyond its walls.

In addition, many Big Data researchers are exploring ways of processing large data sets in memory for the fastest performance. For example, if an organization can cluster five servers that have 256 gigabytes of RAM each and spread a data set across those nodes, it can analyze more than a terabyte of information without the latency of I/O or disk read/writes.

Data Insight Layers

On top of the infrastructure layer are other technologies that give agencies the ability to make sense of Big Data. These are: data processing, integration and management; business intelligence and analytics; decision support, visualization and modeling.

One of the core elements of a Big Data solution is a robust data warehouse. Many agencies have already begun data warehousing initiatives, starting them down the path toward exploiting this robust information resource. A data warehouse is the central repository for an organization's data. It pulls data from disparate operational systems (such as CRM, e-commerce and human resources) and other sources; cleans and normalizes the data; adds metadata to make it easily discoverable by users and computing resources; then stores everything in the data warehouse.

For years, this process has been accomplished using extract, transform and load (ETL) tools, which can be configured to meet an organization's performance requirements. For example, ETL can be used for one-time, on-demand projects, in situations where data doesn't change often or quickly; or ETL can operate in near-real-time as data pours into a data warehouse from operational systems.

Similar in purpose to ETL are open-source tools that have made Big Data solutions available to a great number and variety of organizations. When it comes to tackling Big Data, agencies will want to evaluate MapReduce and Hadoop tools.

MapReduce is a programming framework for processing large datasets across clusters of computers. With MapReduce, a master node breaks down a computational problem and "maps" it to worker nodes. The master node then collects all the answers from the worker nodes and "reduces" them to a single answer. Google uses MapReduce to index the Internet.

Apache Hadoop is an open-source implementation of MapReduce. Hadoop is especially powerful for handling large amounts of unstructured data — the most common type of data in Big Data projects — in part because it can run distributed, parallel processes on commodity servers. Thus, for example, servers that an agency phased out when it moved to a cloud-based email service could be redeployed to handle Big Data processing.

In the same way that companies created commercial versions of Linux, they've also developed commercial versions of Hadoop. IBM InfoSphere BigInsights, EMC Greenplum and Microsoft HDInsight are three Hadoop distributions. Some observers debate whether Hadoop technologies will replace ETL processing, believing that traditional ETL can't keep pace with modern Big Data challenges (especially velocity). Others believe Hadoop is just another way to handle ETL.

Either way, no single solution is the right one for processing and managing Big Data. It doesn't necessarily require a Hadoop distribution, but Hadoop and commodity hardware allow agencies to start small in their Big Data efforts.

Agencies with a significant investment in a relational database management system will likely find that continuing to work with data in a structured format leads to better performance. In those cases, Big Data solutions that integrate the two may make sense.

For example, Microsoft's PolyBase can pull together Big Data from an SQL Server 2012 data warehouse and a Hadoop-compatible HDInsight system. The user builds a typical T-SQL query and PolyBase creates the MapReduce statement, allowing the organization to start using Hadoop without requiring an in-depth understanding of MapReduce.

Many other solutions can process Big Data, some similar to or built on the MapReduce/Hadoop framework. Another popular solution is the NoSQL database. NoSQL databases come in various flavors, but they are not relational databases (nor are they a replacement for SQL databases). They are designed to handle very large data sets and can scale quickly across commodity hardware. Oracle, among others, offers a NoSQL database.

Infographic

Big Data: Access for All

Explore this infographic from Alteryx that offers some tips on how to humanize Big Data: **CDWG.com/planbigdata1**

Analysis and Visualization

Once the information is consolidated in data warehouses, agencies can begin to work with it. There are two main, interrelated ways to do this: analyze and/or present.

Today's data warehouses, MapReduce/Hadoop frameworks and other Big Data solutions offer some analysis capabilities.

But depending on the strategy, other analytics and/or business intelligence solutions may be required, especially in cases where organizations wish to discover information quickly.

Traditional business intelligence tools typically work with structured data, normally an organization's internal data. BI tools may include business performance management, online analytical processing, statistical analysis and other capabilities.

As Big Data grows to encompass unstructured and semistructured data, agencies can apply advanced analytics techniques, such as predictive analytics, data mining and natural language processing, to large data sets. Increasingly, solution providers and researchers are coming up with analytics that take advantage of what is called massively parallel processing.

MPP systems allow organizations to spread Big Data analysis across many computer processors, where it is handled simultaneously (in parallel). MPP enables speedy analysis across vastly larger data sets than traditional analytical solutions. But MPP also usually requires a more complicated architecture for coordinating applications and data sets among processors.

Another class of analytics particularly suited to Big Data is streaming analytics. Traditionally, large data sets are analyzed in batches or "jobs." But in keeping up with the velocity of Big Data, some government groups require high-performance software solutions that can quickly ingest, analyze and correlate data as it arrives from real-time sources, such as social media, voice, video and sensors. Streaming analytics doesn't usually require much storage capacity, but it often exploits parallel processing.

Finally, Big Data projects require tools that present information to users in a clear, accessible way. Modeling, visualization

and decision-support tools take Big Data and its analysis and provide workers a way to explore, ask questions and discover information that supports their missions. Depending on who in an agency will be working with Big Data, visualization tools, such as dashboards, are an important way of hiding the complexity of advanced analytics and enabling more people to take action, thereby enhancing ROI.

Big Data Security

Agencies should understand that Big Data initiatives require a fresh focus on data security. Often, strategies involve pulling together data from separate, siloed sources, most of which, assuming best practices are being followed, are already secure. In other words, agencies may be taking data that is secure and throwing it into a data warehouse that hasn't been adequately locked down.

In general, Big Data security should match the controls of source systems. Data warehouses and Big Data solutions include their own security features. Agencies also can apply security suites to their Big Data resources.

Organizations are advised to perform risk assessments of their data stores and identify where they need to apply security solutions. For example, not only must a data warehouse be secure, but also the processing nodes used to handle the data. Agencies should focus on trusted security measures and apply them to Big Data. Among them:

- Strong passwords (some Big Data systems don't require passwords)
- Role-based access control to Big Data resources
- Network access control
- Operating system hardening of processing nodes
- Encryption of data at rest and in transit