the BIG DATA PLAY BOOK for government

executive

summary

On August 31, 1854, the London district of Soho was struck with a cholera outbreak, and by September 10, the neighborhood had reported over 500 deaths. Curious as to how the outbreak was spreading so quickly, John Snow, often credited as one of the founders of epidemiology, began what would now be considered rudimentary data collection methods in the town.

Snow began collecting information on fatalities and sick citizens and then mapped the location of the data. Through his analysis, Snow deduced that the source of the outbreak was a contaminated water pump, and he convinced officials to replace it. Soon after, the outbreak stopped, and life returned to normal for London citizens. By clustering the data, Snow was able to visualize the outbreak and help officials make an informed decision, saving countless lives.

Today, we have surpassed Snow's primitive — yet at the time, effective — data analysis techniques, and our society is creating more data than Snow could have ever imagined. A recent IDC research report highlights the growth in data:

- From 2005 to 2020, the digital universe will grow from 130 exabytes to 40,000 exabytes, or 40 trillion gigabytes.
- From now until 2020, the digital universe will double every two years.
- The investment in spending on IT hardware, software, services, telecommunications and staff will grow by 40 percent between 2012 and 2020.
- By 2020, as much as 33 percent of the digital universe will contain information that might be valuable if analyzed.

This growth in data represents a remarkable opportunity for government. As public sector professionals, your mission is to learn how to capitalize on your high value and authoritative data. You are responsible for learning what skills, tools and IT you need to transform your agency.

And that's precisely why GovLoop and our industry partners have created our latest guide, The Big Data Playbook for Government.

In this playbook, we provide you with a framework to bring big data to your agency. We break down big data into manageable components to help you understand how to make big data a reality at your agency. Our guide will serve as a roadmap for innovation and provide you with step-by-step instructions to deploy big data at your agency. This playbook will:

- Teach you how to bring big data to your agency.
- Identify which programs to select for a big data initiative.
- Show you what big data is and how to define it for your agency.
- Explore the necessary workforce skills for big data.
- Examine what kind of IT supports big data programs.
- Provide you with worksheets and activities designed to bring big data to your agency and build your big data roadmap.
- Share three government case studies.
- Highlight how industry can support government's big data efforts.

If you've felt like your agency has been sitting on the bench, now's your chance to get in the game. After you go through our playbook, you'll be ready to bring big data to your agency.

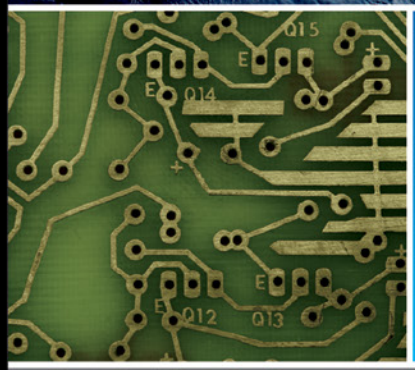
So let's start with the basics on what big data is and how to define it at your agency.

contents

cisco

1

Cisco Unified Computing System



20%

Performance Improvement over Peers When Using Big Data

Industry-Leading Database Performance **34% Faster**

77%

Faster Deployment

Scalable Up to 10,000 Nodes

Find out more at cisco.com/go/govbigdata

Preparing for the Internet of Everything

AN INTERVIEW WITH KAPIL BAKSHI, DISTINGUISHED ARCHITECT, CISCO

According to a recent economic analysis by Cisco, the Internet of Everything (IoE) presents a $4.6 trillion opportunity for public sector organizations over the next decade. IoE is the connection of people, process, data, and things. With IoE, public agencies will benefit from the compounded effect of devices being brought online for analysis. And with more devices becoming connected, organizations will have even more business capabilities and data to explore and analyze, making big data programs more important than ever before.

With IoE, communities can work to tackle some of their most pressing challenges, like improving economic growth, reducing costs, enhancing employee connectivity, and re-imagining citizen experiences. "There are multiple benefits such as operational efficiencies, streamlining mission processes, and creation of cost savings to meet current missions," said Kapil Bakshi, Distinguished Architect for Cisco, Public Sector.

One of the reasons these benefits come about is due to the convergence of cloud, big data analytics and mobility. "All these technologies are working together to give rise to IoE as a next generation platform, and in my mind derive value for enterprises and IT organizations, and that certainly applies to public sector agencies," said Bakshi.

For the public sector, the emerging trend of IoE provides new mission and business capabilties, which previously did not exist or could not be effectively managed. Today, agencies have the chance to create new business capabilties. For example, this could mean changing the way a citizen service is deployed or mission is conducted or a grant is funded, all based on gaining new ways to identify value from IoE.

"Connecting devices so that you can collect unstructured data, and then start to democratize data via a common format and API. IoE presents new capabilities to meet missions, and new capabilities that become available from the benefits of IoE programs," said Bakshi.

But to truly capitalize on IoE, agencies must focus on the convergence of current technology trends. IoE is about the integration of cloud, mobile and Big Data analytics solutions, all working together to create an effective IoE architecture. "In order to realize the IoE and get the value out of it, you have to play to the convergence of the mega trends, which are cloud, big data analytics, and mobility. You should already have a strategy in place for all these key technologies," said Bakshi.

Although all are important mega trends, big data and analytics require special consideration. Today, organizations are collecting more data, and there is a growing need to run analytics to extract actionable value from, mostly unstructured data .

"For IoE you need to have an infrastructure that has both edge and core analytics capability," advised Bakshi. "Your infrastructure must not only be able to provide analytics from streaming and real time at the edge of the network, but also summarize and it send back to the core for batch and historical analytics."

There have already been dozens of examples of use cases in the public sector of IoE. From first responders to war fighter to to law enforcement officials to educators to scientific missions, many have used IoE to re-imagine how they accomplish their mission.

"Healthcare is one area where you see a lot of innovation happening, and a lot of development for new business models, with a lot of new mission capabilities being developed from IoE. This includes IoE applications from patient care, clinical trials and the pharmaceutical industry," said Bakshi.

Another key area of innovation is Cyber security. In the IOE architecture, cyber security would play a very important role as the number of connected devices grow exponentially and hence the threat perimeter grows accordingly. Moreover, cybersecurity is becoming a key use case for IoE and Analytics, as one uses cyber security forensic, anomaly detection and mitigation approaches.

In order to get started with IoE, there are some preliminary steps that agencies must take. One is that agencies should focus on defining a specific mission or business problem they are attempting to solve.

"With the use case and end goal in mind, and the mission problem you're trying to solve, you should see how technology trends converge to give you a starting point for your IoE projects," said Bakshi.

Once the desired program is selected, organizations must then create a technical and business strategy. Cisco can then help agencies deploy IoE and identify new ways to find value from their data.

"In not just the technology aspect of it, but also the mission strategy part of it, Cisco can help you create that strategy for the specific use case or the specific business problem that you're trying to solve," said Bakshi.

how to think about BIG DATA at your agency

LEARNING OBJECTIVE

After going through this section, you and your team will have a better understanding of what big data means for your agency. You'll also be able to clearly define the problem you are trying to solve with a big data initiative and be able to navigate some common obstacles.

What is big data?

By conducting a quick search on the Internet, you can get hundreds of conflicting definitions of big data. The problem? Big data means something different to everyone. For instance, the way that big data is defined by the Department of Commerce is different than the city of Syracuse. Their definitions vary in applications, use cases, and the size and scope of projects.

But there is still a commonality across different states and levels government: Big data is leveraging emerging IT to unlock the power of your data and drive improved decision making for business operations. So the focus should not necessarily be on scale or size of your data but rather about the opportunity that big data presents to transform your business processes.

Quite often, you'll see big data defined using the four Vs:

- **Volume:** The quantity of data that your agency collects.
- **Velocity:** The speed at which data is created.
- **Variety:** The various data types that your agency has access to.
- **Veracity:** The authoritative nature of government data.

But even given this working definition, it's important to think about big data not so much as just a technology, but as a phenomenon or event. Think of it this way: Your agency is creating troves of data daily, and this data represents a new way to look at business processes, services and constituent services. So big data is not solely linked to just technology — it's also part of a way of thinking. Big data has become both an art and a science. That's why we want it to become part of your organization's ethos, so you can use it to create data-driven solutions for your agency.

But (sorry to do this to you) we've got a big caveat here: We called our definition a working definition for a reason, and that's because, in time, your agency's definition of big data will evolve. You'll grow in size of your data. Your storage needs will change. And above all, you'll mature in the complexity of your big data program. So as you go through this playbook, we want to make sure that you are armed with the right knowledge to rise to the challenge, seizing the opportunity that big data presents.

For every agency, the big data journey starts by defining the problem you need to solve. Next up, we give you some insights on how to get started with your big data program.

What problem are you trying to solve?

Now that we have run through an introduction of what big data is, how do you get started? We know this sounds a bit basic, but it's so important to discuss. The key is that you must be able to answer the question: What problem are you trying to solve with big data?

You've already seen that there are dozens of ways your agency can use big data, and the applications can seem endless. But what we've seen is that organizations that pick one core mission problem, and then gradually build and grow programs based on lessons learned, have better success with big data.

So how do you begin and pick the right program? It's important to think about building your roadmap and starting small on some projects that will have quick wins.

If this is your first big data project, you really need to be smart about the project you're picking, as you want to start off on the right foot. And, if you've done a big data project before, you want to guarantee continued success. So like any strategic planning exercise, to select the right big data program, you must ask the right questions. Here are ten to get you started:

1. What problem are we trying to solve? How will data help?
2. What outcomes do we want to achieve?
3. How will big data work to meet our mission needs?
4. What kinds of data do we need access to?
5. Who are the main stakeholders and how do we engage them?
6. How are we going to track, assess and monitor progress?
7. Can we pilot a few programs and start small? What can we learn from starting small and building out?
8. Do we have the right workforce in place?
9. Have we received buy-in from leadership and across teams?
10. Are we delivering on the needs of our users? How do we know?

Now, after you've answered these ten questions, you're ready to start diving into big data. Before you advance on, take a laser-like focus and address the following three areas:

DEFINE YOUR OWN MEANING OF BIG DATA

Everyone is going to define big data differently. Start by understanding what big data means for your agency and placing the Vs of big data into the right context for your agency.

START SMALL

You're going to want to be sure to start small. Running a few pilots around big data can't hurt; this will help you get a better understanding of the lay of the land and what you can improve with data.

SET CLEAR AND MEASURABLE OUTCOMES

Be sure that you are measuring success and thinking critically about what your success metrics will be. You must have clear and actionable goals that you want to achieve with your big data program, like catching fraudulent activity 90 percent of the time, automating processes to improve workforce efficiency by 20 percent, or increasing speed of FOIA requests by 35 percent.

Now, our astute readers (that's you) might have noticed we left out one burning question: What challenges might our agency face in using or deploying big data? This is a very important question to address, but we wanted to wait a bit to talk challenges. For success with big data, you must start by thinking about goals and objectives, not roadblocks and barriers.

But of course you'll face challenges. So once you've clearly defined your objectives and what you want to accomplish, now is the time to think about potential barriers. That's why our next section walks you through some common challenges.

Navigating Common Obstacles

Like with any kind of IT deployment, you're going to run into some challenges with big data. We surveyed the GovLoop community to learn their top challenges and have synthesized them here, along with some best practices for you to follow.

As you start your big data journey, you can't forget that big data is not necessarily going to be an easy road to travel down. So it's critical to stay focused on the power of big data and the opportunity it presents for your agency. In the meantime, here are a few common challenges — and ways to work through them.

CHALLENGE #1: A LACK OF LEADERSHIP SUPPORT

So you're passionate about big data. You've been a champion of it. But you're struggling to get the message through to senior support.

Big data can seem daunting to leaders, and no matter how great the benefits might be, your agency's senior leadership must take a thorough look at the program and make sure it's fitting into your organization's strategic objectives. And that's a fair and perfectly reasonable approach to take prior to launching an initiative.

But you need that buy-in for your big data programs. So what are some go-to plays to gain the support you need? Here are a few techniques to deploy:

PROVE THE BUSINESS VALUE

With limited resources, decisions that are made in the public sector must revolve around obtaining new business value. If your organization has core metrics you are trying to achieve, or deliverables you must meet, show how big data can help and exceed those goals.

SHARE USE CASES OR EXAMPLES

Another strategy is to talk about success studies from your peers. The more you can show achievements from other government agencies and explain how you can replicate that success, the more likely you are to gain buy-in. Again, you must be able to provide a clear problem that you'll be solving by leveraging data.

CREATE COMMUNITIES OF PRACTICE

Struggling to get support? Consider connecting with like-minded professionals. Big data requires the sharing of knowledge, best practices and relevant case studies. By building a community of practice within your agency, you can help build support across your department, encourage knowledge sharing, and start to build a data culture. Eventually, senior leaders will take notice and want to be part of the positive energy. Our NASA case study on page 9 is a good example of how important it is to build consensus when initiating your big data program.

CHALLENGE #2: A SLOW ACQUISITION PROCESS

We hear this all the time at GovLoop: Government's acquisition process is too slow, and it doesn't allow you to quickly procure the tools you need. Unfortunately, in a lot of cases, government's hands are tied. You need to adopt new technology, but the process is slow and you have limited funding. To work around this, contemplate the following ideas:

CONSIDER YOUR ENTIRE BIG DATA INFRASTRUCTURE DESIGN

You need to consider everything about big data, from your servers, cloud computing, compute, storage and management needs. Think about this as your "big data ecosystem," and how everything works seamlessly together with little downtime. Once this is done, maybe there are some quick wins you can earn, and use them as ways to drive innovation, while using your agency's existing technology.

PLAN FOR THE FUTURE WHEN PURCHASING

Any IT purchase is a long-term investment. If you're upgrading from legacy systems, be sure that your solution will be able to scale up to meet future needs. This way you can be prepared as your agency continues to collect and manage even more data.

EXPLORE SHARED SERVICES MODELS OR LEVERAGE EXISTING IT

Don't have access to the cloud or servers? Maybe there is a shared IT model you can use across your agency, or there are easier ways to get access to technology. It's also possible that you could re-engineer some existing IT to meet your infrastructure needs.

CHALLENGE #3: LACK OF CLARITY ON DATA NEEDS

One of the biggest challenges for big data is just answering these questions: What data does our organization have? What data will be useful?

The reality is that so much data is collected, stored and managed — often across multiple departments and agencies — that it's a common struggle to find value with all this information. So where do you begin? Here are a few ways to start:

CONDUCT A DATA AUDIT

A data audit is essential to understand what your current data landscape looks like. A data audit will help you realize what data is missing, what's needed, and where it is located. This will be an essential step to your big data program. This step will be significantly more time consuming and challenging than you think, so be sure to budget adequate resources.

DON'T REINVENT THE WHEEL

As you are going through your big data audit, you may see data redundancies or duplicate work. Try to avoid reinventing the wheel, and work across teams to understand their data and information. This can help be more efficient and effective as you embark on your big data program.

PRIORITIZE DATA

Prioritizing your data is a critical step. You must know which data is the highest value to your organization. This will dictate the proper steps to make and correct ways to migrate essential data. It will also allow you to put in place the right security requirements around your data.

CHALLENGE #4: MIGRATING DATA & APPLICATIONS EFFICIENTLY

Once you're ready to do a big data project, there may be a lot of data that has to be migrated. Unfortunately, this is not just as simple as "copy and paste" from one server to the next. You need to be sure that as data is migrated, it will not interrupt peak performance times or user needs. Once the migration is complete, you must assure that all systems are running efficiently for users. So where to start?

DEFINE PERSONAS TO MANAGE DATA

Every employee will need different kinds of accessibility, so make sure that your data system maps to these needs and is not providing unauthorized access information.

GOVERNANCE IS ESSENTIAL

Having a governance policy is essential to your big data project. You need to be able to manage your data, and users need to know what is an acceptable usage of information. Be sure that your governance policy is well defined and supported by the team.

COLLABORATE ACROSS TEAMS TO AVOID DOWNTIME

When you're ready to move data, there will potentially be some downtime. You must be sure to connect with your team, so you can be assured that when migrating data, it is done at a convenient time for users, which will avoid an interruption of work.

CHALLENGE #5: CULTURAL ROADBLOCKS

Big data is going to impact a lot of processes at your agency — and potentially even your culture. There will be changes in workflows, accessing work, sharing data and collaboration. So to implement new processes, you need to be smart in how you approach change management. Here are some ways to break through cultural barriers:

BUILD A CULTURE OF TRUST

Yes, this is easier said than done, but having an environment where people want to share data, work across organizational boundaries, and care about elevating the mission of the agency (and not just one department) is imperative to fully leverage the opportunity big data presents.

ALLOW FOR RISK, AND EMBRACE FAILURES

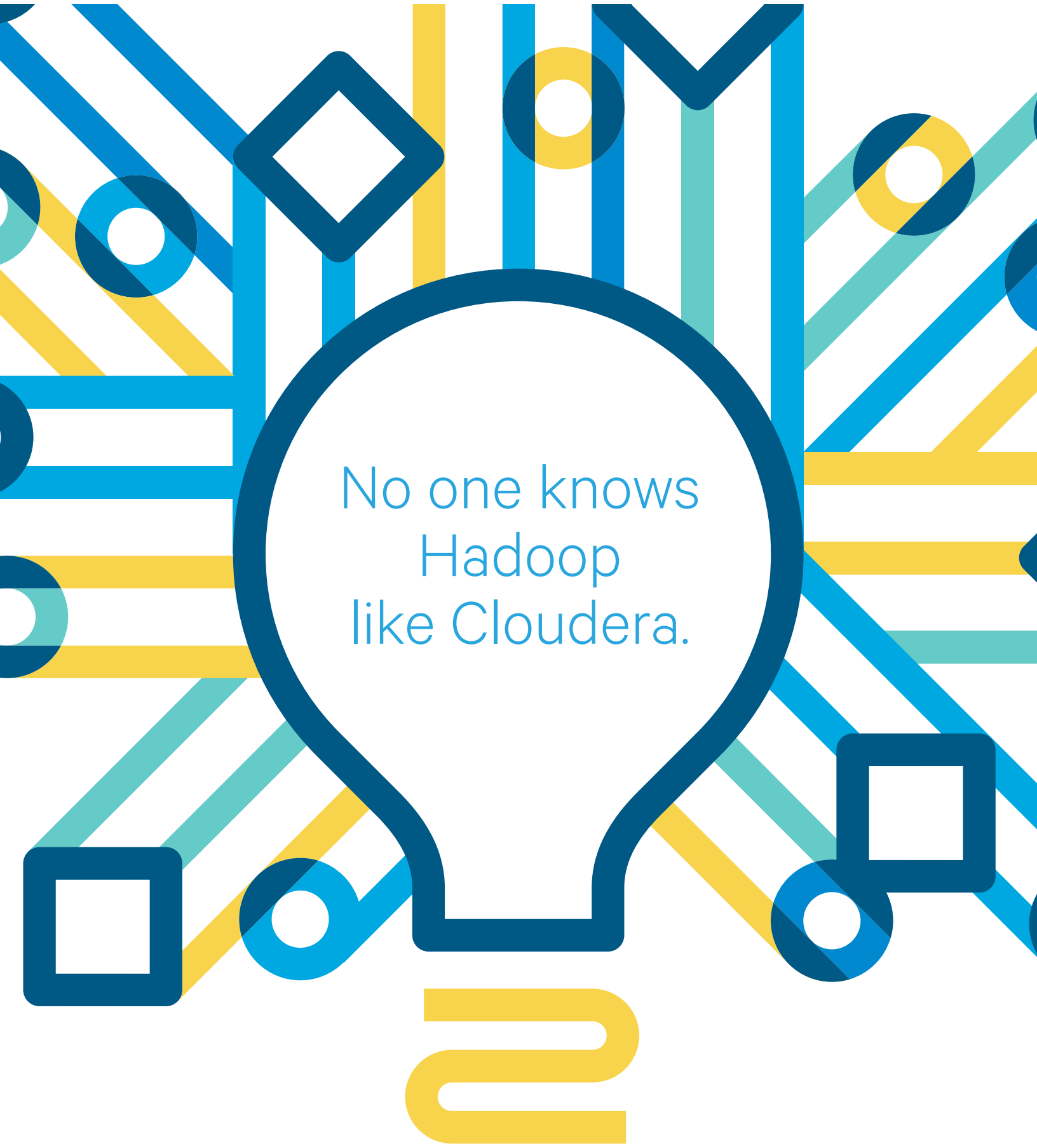
There's a great Bob Dylan quote: "There's no success like failure, and failure is no success at all." In other words, your organization needs to take risks, and if projects do not turn out as intended, embrace it as a learning experience. Take the lesson learned and move forward with a renewed vision and enthusiasm for the next big data project.

FOCUS ON EDUCATION AND TRAINING

If you're running into roadblocks on culture, consider offering some training and education to your team on big data. This will help them clearly see the impact of the program and how big data can improve organizational efficiency.

Build your Big Data Roadmap

Looking to build your big data roadmap? We're here to help. Stay tuned throughout our playbook, as we'll remind you of important activities and templates to check out in the appendix. For our first set of recommendations, check out Worksheet 1: Creating Your Big Data Statement of Purpose.



cloudera

Jumpstarting Your Big Data Campaign

AN INTERVIEW WITH WEBSTER MUDGE, SENIOR DIRECTOR OF TECHNOLOGY SOLUTIONS, CLOUDERA

One of the challenges for any IT program is soliciting support from leaders and gaining executive buy-in. Although technology is important, big data starts with developing the right partnerships internally to drive success. Webster Mudge, Senior Director of Technology Solutions at Cloudera, recently spoke with GovLoop about how organizations should start their big data journey and the important steps along the way.

"One of most important elements, and a really critical function of creating a big data program is that you have to develop confidence and sponsorship from leadership," said Mudge. "You need leadership to foster curiosity and sharing. And that is really a prerogative if you want to go and make big data a critical asset. Yet, there is a level of risk that the sponsor has to take on – an executive, or any business mission leader, must be focused on change and also be willing to take a risk to navigate change to ensure the success of the overall program."

In other words, organizations that are serious about big data must understand that it's not solely about technology; it's also about the culture change. Another critical aspect to navigating and fostering culture change is building advocacy within the agency. As Mudge reminds us, sharing stories of big data success can often accomplish this critical step.

"You have to build success stories from the point of view of your organization in order to push this forward," Mudge advised. "Begin with something straightforward. Start with something that is demonstrative and will add value, that will also showcase cost savings and make processes more efficient."

There are also some preliminary IT steps that an agency should take to begin to leverage big data. A good starting point on this journey is often just basic, single data set analysis. This means looking at a individual sources of data, cleaning the data, and starting to process the data separately, then making them available in various tools and environments that are found in Hadoop and the Hadoop ecosystem. These could be tools like search or business intelligence solutions.

This gets employees comfortable with data, and helps to create a data-driven organization. In many cases, this step is where people can begin to witness efficiencies from better use of data, the cost efficiencies and option value of working with data in this manner, and helps show the importance of having a strong IT foundation to build on success.

The next phase is expands the scope from single data set analysis to multiple data set analysis. This means beginning to understand how individual, often disparate, data sets are correlated, and running analysis on this more complex data and developing in-depth trends.

Once agencies have started to build the foundations, identify some success, and gain some quick wins, Mudge said, they should then look to explore building an ROI model, which will help support emerging applications of big data at the agency and provide the necessary feedback and information for the program sponsor. When some initial success has been achieved, especially with cost savings from efficiencies, savings can then be re-invested to support the core mission of the agency, and help drive more advanced data programs. This activity can be critical as a program moves into multiple data set analysis, says Mudge, because often multiple data sets means multiple sponsors, and ROI and feedback on prior successes can better champion the program with new leadership.

The final two phases of a big data journey include the idea of predictive and operational analytics. In most cases, predictive analytics is applying statistical models to predict future behavior, which will drive new value from data. With operational analytics, agencies can begin to deliver near real-time value, within and throughout the entire mission lifecycle.

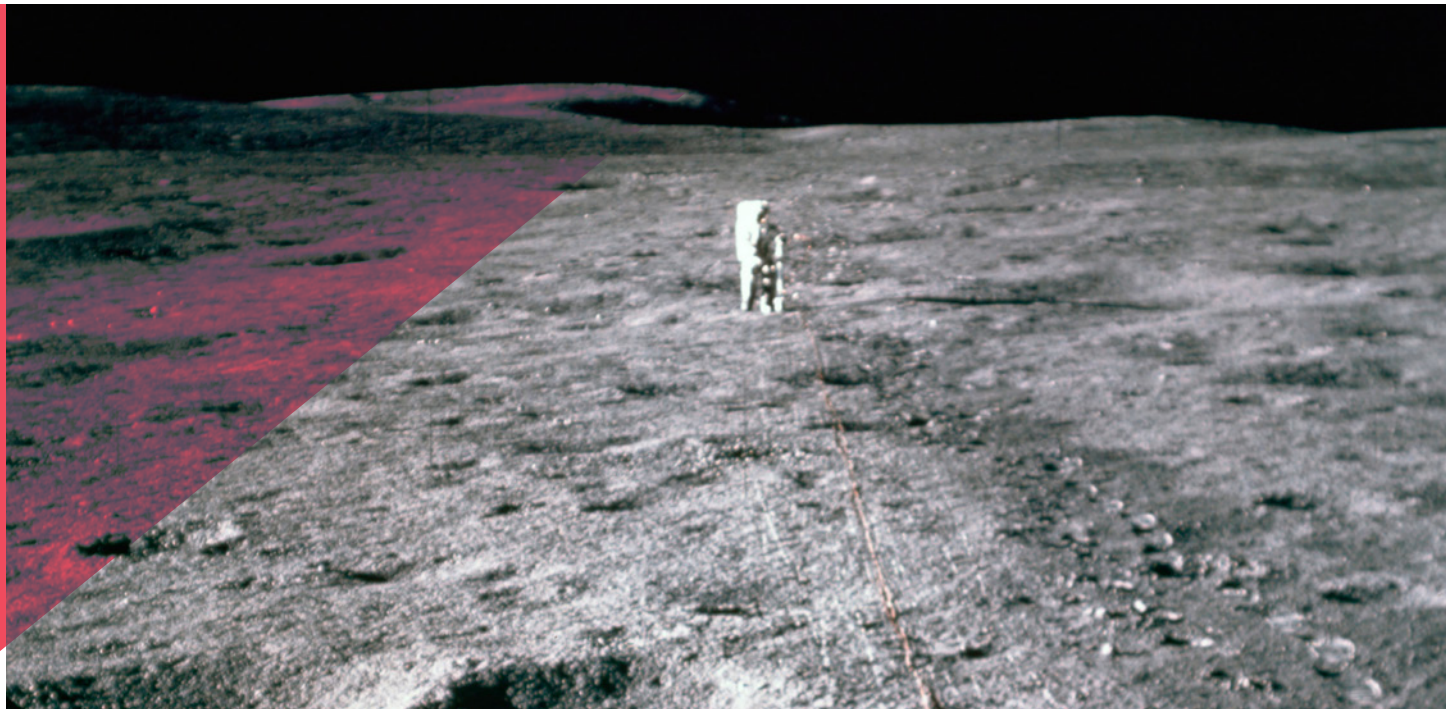
"At this point in the journey, analytics becomes pervasive throughout the entire value chain of data, not only embedded throughout the process flow of ingest and discovery and processing, but also embedded throughout the value chain of the mission workflow and applications," said Mudge.

With these analytic solutions, organizations' workflows change, impacting how the back office and field systems work. This allows organizations to gain both a historical view of the data and create real-time models, leveraging all data within the agency. And this shift and activity fosters culture change within the agency, which is a primary reason for developing strong leadership for the program.

"The way Cloudera looks at an enterprise data hub (EDH) architecture is that it is a wealth of potential for your agency," said Mudge. "Yet, an EDH is not going to be institutionalized or come to fruition unless you have a plan on the business side as well as the technical side, so the two tracks have to happen simultaneously. As an enabler for the technology, we have to spend time on fostering and building that business process transformation story."

SPACE: the next data frontier

On May 5, 1961, Alan Shepard became the first American to enter space. Shepard instantly became a household name, along with the federal agency that sent him there: the National Aeronautics and Space Administration (NASA). Riding NASA's Freedom 7 spacecraft, Shepard left Earth's atmosphere and made history.



Shepard's trip was a monumental achievement for NASA. The agency was formed only three years earlier and was in a tight space race against the Soviets. But as history shows, NASA prevailed in the space race and to this day continues to do remarkable work and push the limits of science and the exploration of space.

Ever since NASA was founded, data has been at the core of its mission, and today, the agency is a leader in big data technology. Whether it's the programs designed to send images back to Earth from satellites, or navigate the terrain of Mars, the agency is collecting more data than ever before. And not only are they collecting, managing and storing data, they are working hard to make it actionable and available to scientists, staff, researchers and citizens.

GovLoop recently interviewed Nicholas Skytland, NASA's Big Data Evangelist. Skytland shared current applications of big data and some lessons learned from working at NASA.

"In my role, I am educating NASA on why data is so important and how it is the currency of the information age. Today, it's more important than ever before for NASA to be looking at new ways to leverage big data," said Skytland.

Although the agency is collecting more data than ever before, Skytland reminds us that not all data is created equal. Prior to being made public, some NASA data must be vetted, cleaned and assured that no security risks are hidden with the data.

"[Data.nasa.gov] is just the tip of the iceberg of the data that NASA has internally. A lot of the efforts we are doing right now are to work with different divisions and projects within our agency," said Skytland.

Skytland also walked through some of the challenges that he has faced using big data at the agency. "I like to break down our challenges as the management and processing of our data, storing our data, archiving our data, distributing our data, making it more accessible, and finally the analysis and visualization of data. Closely related is the idea of commercial cloud computing resources and how we as an agency work through that as well," said Skytland.

While working with big data teams, Skytland has found that many employees have a high level of expertise in a specific area but struggle on what to do with their data. For instance, someone might be great with data storage on a cloud platform but struggle with what to actually do with the data. "In this instance, we can work with them on the analytics and virtualization part of big data and show them the possibilities if they unlocked data using an API," said Skytland.

NASA faces similar challenges as other federal agencies, like obtaining the right technical requirements and fulfilling business needs through big data. Some of the challenges that Skytland identified were:

- How do you make a business case within federal agencies for investing intentionally in big data solutions?
- How do you communicate [big data value] with senior leadership and management?
- How do you inform science and program managers about the right way to go forward, when they have so many different things to do?

"Hopefully [we can help adopt] big services within the agency, so that people can stop worrying about big data and start worrying about achieving their missions," said Skytland. To help understand how to deploy big data services, like assisting with data accessibility, storage and computing power, Skytland identified some best practices from his work at NASA.

"We listen to a department's problems and understand what it is they are trying to solve and what it is they are trying to address. We really try to be intentional about understanding challenges, not prescribing a solution up front. Once we understand that, we dive into the particular data they have in that context. So it always starts with listening," said Skytland.

"The end goal is that we don't want anyone to worry about big data, which is essentially any amount of data that you can't handle," Skytland continued. "We want to focus on exploration, reaching new heights, or unveiling the unknown that benefits all of humankind."

As a final piece of advice, Skytland says you must never underestimate the power and potential of data. "There is so much potential that is often untapped in data, that if we can figure out ways to give more access to data, I think we will see a huge benefit in that."

Building Your BIG DATA All-Star Team

Your learning objective:
This section will help you understand what people and skills you need on a successful big data team. You may not be able to hire additional staff, but you can certainly assess your workforce and understand any skills gaps and how to fill them.

In order to take advantage of big data, you're going to need a highly skilled staff. But what do they look like? What talents do they need? Who are they and where do you find them? Everybody's big data team is going to look a little different, but here are some of the core team members you're going to need:

We know that hiring or filling all the necessary positions for a big data program can be a challenge, especially given the current state of budgets in the public sector.

So be sure to check out Worksheet 2, which provides a worksheet on how to prepare your workforce for big data and Worksheet 3: Your Big Data Workforce Planning Sheet, both in the back of this report.

These activities can help you spot gaps and the needed skills at your agency for a big data initiative. And later on, we learn more about how to the prepare the big data workforce from the National Institute of Standards and Technology (NIST).

THE CHIEF DATA OFFICER

A Chief Data Officer (CDO) is playing an increasingly important role for government. The CDO can help an organization formalize their data strategy. The CDO serves as an executive position, which helps the agency transform how data is collected, stored and managed. The CDO should:

- Operate as a strong team builder, have exceptional interpersonal skills, and can speak both IT and business language effectively.
- Provide strategic direction for data applications across the agency.
- Have extensive expertise in market trends on data and analytics.

THE DATA SCIENTIST

A data scientist serves a crucial role on your big data team. The data scientist must be able to understand how to derive value from multiple data sets and clean data sets for end users. In short, the ideal data scientist should have the following skills:

- Knowledge to query databases and conduct statistical analysis.
- Ability to create, manage and operate databases for big data programs.
- A thorough understanding of business strategy and the link to organizational data.

THE PROGRAMMER/ENGINEER

A big data programmer works closely on parsing, managing and analyzing large sets of data, which ultimately can make data actionable. The big data programmer should be able to build prototypes of what the big data solution will look like and work closely with other teams to assess effectiveness. They should be well versed in a variety of different big data solutions and tools. Their skills include:

- Expertise in big data technologies (Hadoop, for example).
- Ability to document, track and manage big data programs.
- An understanding of cloud computing technology and how to leverage it for big data programs.

COMMUNICATIONS TEAM

The communications team will be critical to your big project, as they will be the ones promoting it to citizens and internal stakeholders and receiving feedback on the initiative. They can help your team market the solution, gain buy-in and understand some challenges for their perspective. They will:

- Communicate to external and internal stakeholders about program success.
- Promote and advertise the big data program across your team.
- Work across teams to learn about initiatives and impact on organization.

CITIZENS

Ultimately, your goal is to meet the objectives of your mission and provide better solutions for citizens. In some cases, citizens will play a critical role in your big data program, and you need to know what they like, don't like and what services they believe are necessary. Someone at your agency must be able to:

- Gather feedback on citizen and user needs.
- Assist citizens as they develop applications and tools from your open data programs.
- Encourage your agency to adopt citizen science initiatives for additional data collection.

INDUSTRY PARTNERS

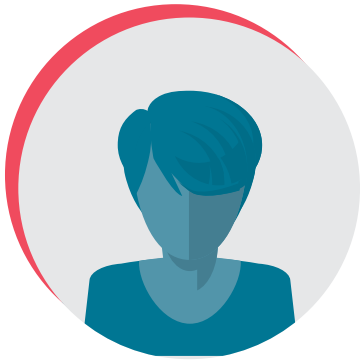
Industry plays a critical role in your big data program. Industry partners help you obtain the necessary solutions and support. They can provide you with the latest industry solutions and make sure that your software and solutions are updated and secure to reduce security risks. Vendors can provide:

- World-class IT and infrastructure to support government agencies.
- Cutting edge technology to push big data programs further.
- Consulting support and advisement on technology needs.

THE BIG DATA ANALYST

The big data analyst should enjoy looking for the needle in the haystack and understand how data impacts policy decisions. Big data analysts could have degrees in public administration, statistics, computer science or a variety of disciplines. The skill sets include:

- Articulating data and findings across teams and to senior management.
- Understanding gaps in knowledge and what data might be needed.
- Providing insights on potential impacts from data-driven policies.



THE BIG DATA PROJECT MANAGER

The big data project manager is someone who can bring all the teams together and keep the project on scope and within budget. This will allow production teams to focus on building solutions and meeting objectives. Just like any other project management role, this role's responsibilities include:

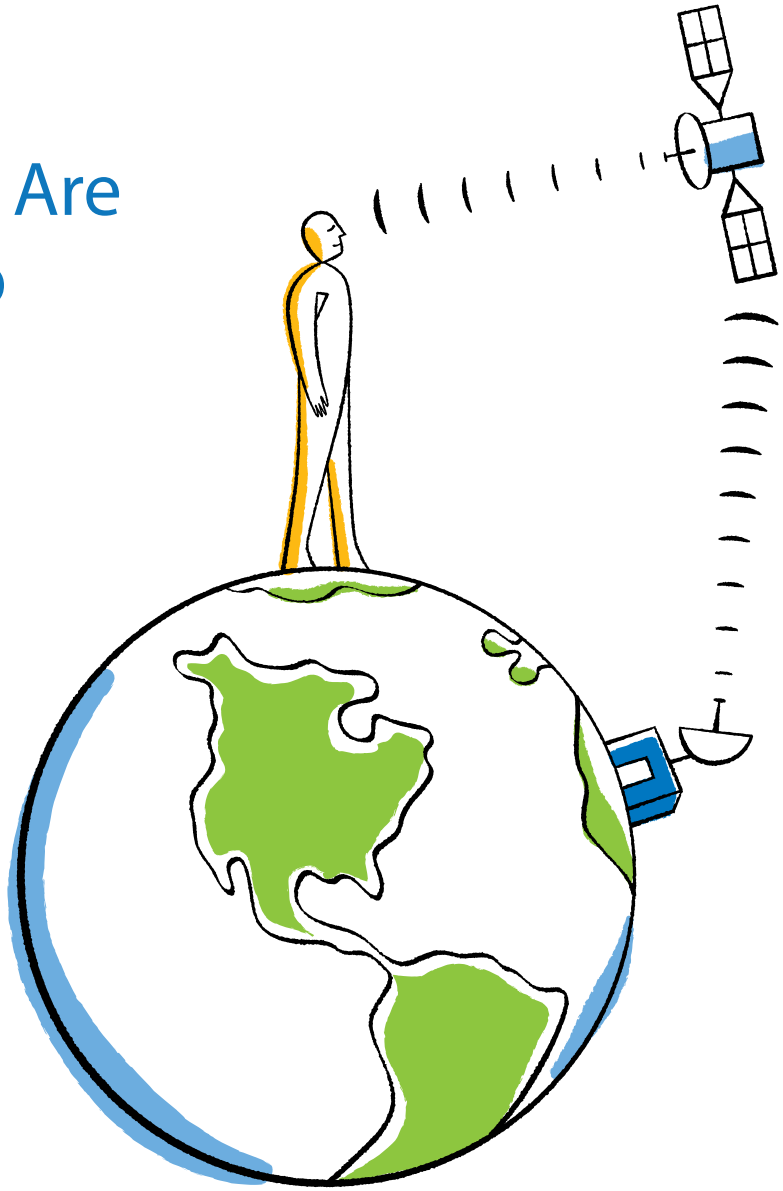
- Serving as project lead, making sure projects stay on time and on budget.
- Working across teams as a liaison, speak IT and business language.
- Understanding of basic big data technologies and programming languages.

THE BIG DATA VISUALIZER

A big data visualizer is someone who can take raw data and turn it into beautiful graphics, illustrating your data. This role may meld into other positions on your team, but some necessary skills include:

- Understanding of programming languages and graphic design techniques.
- Extracting key data and working across teams to understand visualizations.
- Deep understanding of agency mission and importance of data.

Critical Missions Are Built On NetApp



NetApp's Big Data solutions deliver high performance computing, full motion video (FMV) and intelligence, surveillance and reconnaissance (ISR) capabilities to support national safety, defense and intelligence missions.

To learn more about how NetApp can solve your big data challenges, visit us online at www.netapp.com.

Understanding the Universal Data Platform

AN INTERVIEW WITH DR. GREG GARDNER, CHIEF ARCHITECT, GOVERNMENT AND DEFENSE SOLUTIONS, NETAPP

With government agencies creating more data than ever before, many organizations are leveraging emerging technology to create new kinds of data infrastructures. GovLoop sat down with Dr. Greg Gardner, Chief Architect, Government and Defense Solutions at NetApp, to discuss new data platforms for cloud, mobile and big data. These solutions can enable the public sector to make sense of the huge amounts of information being created today.

"We are seeing a combination of on-premise and cloud computing in an increasing number of organizations to address their data needs," said Dr. Gardner. "This combination of on-premise and cloud computing is what we call a hybrid IT infrastructure – where some IT is in the cloud and data is stored primarily on-premise."

NetApp's distribution partner, Arrow ECS, enables agencies to manage their hybrid cloud environment from a single pane of glass with a robust portfolio of cloud services and products and a management platform called ArrowSphere.

"An agency's IT infrastructure has to be replicated and made available to cloud computing capabilities to create one common data infrastructure for compute," said Dr. Gardner. "So regardless of whether you do compute on-premise or in the cloud as a service, people have access to the same data and information."

The hybrid IT infrastructure creates a data-centric model, which is essential for the success of a big data program. The data-centric model replicates data, so that regardless of location, people are accessing the same data, whether it is on-premise or in a cloud service. To create this kind of model, many organizations are looking to build a universal data platform, providing access to information regardless of where data is hosted. In creating this universal data platform, there are some common features that must be included to ensure a consistent service quality across all cloud and data center assets.

These features usually include secure multitenancy – the ability to host multiple workloads and data belonging to multiple organizations or functions all at the same level of classification, without degrading any services or data boundaries.

Another element is the ability to pool virtual resources, the ability to abstract hardware resources, irrespective of their location or types, and provide efficient data transport and management features.

For government agencies, it is essential when moving and storing data that information is immediately available whether an employee is operating in a cloud infrastructure or an IT on-premise infrastructure.

"If you just have all the data in one place, you are limited as to how you can use that data," said Dr. Gardner. "We talk about data like it is roaming charges on a cell phone plan; you can store it very cheaply for pennies for gigabyte per month, but if you go to move the data, it gets very expensive. A universal data infrastructure makes data available regardless of when, where, or how people are using it. We believe that is really key."

In addition to this hybrid model, agencies are also exploring the benefits of a converged infrastructure. A converged infrastructure refers to a package of compute, routing, switching, storage, and virtualization so that all these functions are provided together in one package, ideally supported by one agent.

For instance, a converged infrastructure will include varying amounts of compute or storage, whatever is required to meet the mission at hand. All of these are preconfigured packages with one point of contact for help desk services. Having server, switching, compute, storage and virtualization in one package makes the infrastructure easier to manage, controls costs, and enables scale up or out to meet user demands.

Dr. Gardner hypothesized a typical use case that shows the power of a converged infrastructure and leveraging a hybrid IT model. Take for example a county government with an aging data center that needs to upgrade its infrastructure. The county has legacy IT systems, a limited workforce, budget limitations and regulatory requirements. Those factors combine to constrain the way they do business, and create challenges to innovation in service delivery. To work around these obstacles, the agency could maintain control of their data on-premise, replicate the data to a co-location facility and then, using high bandwidth pipes, expose it to the cloud. In doing so, the county could leverage low-cost, competitive compute capabilities from a number of vendors while maintaining control of their data. The cost savings and significantly increased capabilities of this approach offer real advantages to the county.

"There are certainly hurdles in security, data access and management and fiscal constraints in these approaches," added Dr. Gardner. "But we believe the challenges can be overcome with a hybrid IT approach that embraces cloud computing as well as on-premise-based data management. We strongly believe the idea of converged infrastructure and a universal data platform. This model can be realized with the appropriate architecture and working with partners like Arrow to get public and private IT organizations that kind of solution."

closing the gap: TRAINING public sector data scientists

"Big Data has spawned the discipline of 'data science' to derive knowledge from very large-scale complex archival and streaming data from almost every domain. Data science is driving crucial decision making in many sectors,"

ASHIT TALUKDER,
Division Chief, Information Access Division at NIST

The National Institute of Standards and Technology has been one of the leaders in helping government agencies understand how to deploy big data initiatives. One program in particular, the NIST Data Science Program, helps teams assess their big data workforce needs.

"Our vision is to lead an interdisciplinary multi-sector data science program initiative focused on enabling the understanding of data science approaches and driving advancements through benchmarking, reference data, challenge problems and rigorous measurements," explained Ashit Talukder, Division Chief, Information Access Division at NIST. Talukder said that the NIST program works to achieve their goals by:

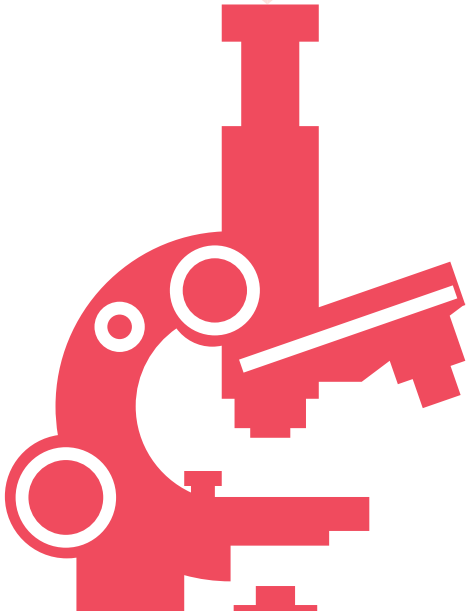
- Working with the data science community to address technology and measurement barriers.
- Fostering advances in data science through:
  - Advancement of **rigorous measurement technique.**
  - Development of **reference frameworks and reference data sets.**
  - Development of **open challenge problems** on **use cases addressing compelling classes of technology challenges.**
  - Community collaboration **(engaging stakeholders from all sectors).**
- Helping stakeholders understand the state-of-the-art technologies, facilitate collaboration, direct comparison of approaches, and create basis for future standards.
- Enabling open dialogue among multi-stakeholder communities to discuss common interests, challenges, problems, solutions, and future directions in big data and data science.

"In summary, the objective of the NIST Data Science Program is to improve the reliability, resilience, access, robustness, accuracy, generalizability, security, usability, and performance of data-driven discovery and decisions through measurements and standards," said Talukder.

The NIST program comes at an important time, as government continues to deal with larger and more complex data. To capitalize on the promise of big data, agencies must have the skill sets to properly capture, store and extract meaning from the information.

But agencies face many challenges along the way. "While it is evident that big data is already resulting in improved data usage and has an increased role in decision making across multiple domains, the rapid growth of big data technologies by multiple stakeholders is resulting in solutions that often cannot be adequately measured and characterized or interoperate with other solutions," said Talukder. He also identified the following challenges:

- Lack of sufficient big data interoperability and data standards.
- Lack of efficient interoperable frameworks for big data analytics.
- Need for standardized interfaces and interoperable interfaces/APIs.
- Need for reference frameworks for big data solutions and systems.



Talukder revealed another area that needs to be addressed with big data: performance measurement and characterization of big data and data science solutions. These challenges include:

- Limited coordination between big data experts and scientists across sectors and domains.
- Limited analytics solutions that work robustly with multimodal and heterogeneous data types.
- Lack of understanding of what works (and does not work).
- Lack of objective understanding of the foundational gaps in big data science.
- Lack of accepted evaluation methods, tools and reference data in big data science.
- Lack of understanding regarding usability of big data systems and tools.
- Limited understanding of uncertainty propagation in big data systems and how the quality and context of the input data affect the resulting discoveries and derived conclusions.
- A lack of multidimensional benchmarks that can be applied to the analytics tools and processes.
- A way of evaluating which components are best suited for specific families of tasks.

To assist agencies and develop big data standards, interoperability and reference frameworks, NIST has created a public working group dedicated to big data. And they have also worked to advance the study of data science.

"To help advance data science in multiple sectors through measurements, benchmarking and reference data, NIST hosted a Data Science Symposium in March 2014 (http://www.nist.gov/itl/iad/data-science-symposium-2014.cfm). The intention of the symposium was to bring together a diverse community of data science practitioners and technical stakeholders to discuss challenges and opportunities with regard to measurement science for data science technologies and methods," said Talukder.

BIG DATA

Navigating the People, Process & Technology of Big Data

AN INTERVIEW WITH MARK A. JOHNSON, DIRECTOR, PUBLIC SECTOR BIG DATA PROGRAM, ORACLE

Although big data has the potential to radically redefine government operations, organizations still face significant obstacles such as training, integrating disparate data sources, and keeping costs low. As is the case with any kind of IT initiative, to succeed in these initiatives, organizations should look to integrate big data into already existing standards and procedures used at the agency.

"Organizations really need to simplify the process and use the skillsets that people have, advised Mark A. Johnson, Director of the Public Sector Big Data Program at Oracle. "Training and hiring people is difficult and expensive, but adopting interfaces that you have and leveraging existing skillsets is relatively inexpensive. "There may be some upfront costs at times and new IT tools, but in the end, the tools are going to allow data to be leveraged much faster. The idea is that you don't want to train staff on new IT – you want them to use existing tools to leverage big data."

Users should have easy access to data and information anywhere in their agency. But the challenge then becomes knowing where to store data, since different data should be stored in different ways. Additionally, to improve data utility, agencies should strive to minimize data movement between different data stores and locations.

"Relational data still belongs in the relational store, but unstructured data may belong in a Hadoop Distributed File System (HDFS), or in a NoSQL database," said Johnson. However, no matter where the data is stored, analytics should work across all data, and provide an analyst access to any information that is needed.

"Different analysts may have very different needs as to the kind of analysis they want to conduct," explained Johnson. "So agencies need to be agnostic at the data store level of who's getting data and how they want to receive access. People want to analyze data, check it out, look at it and use it in different ways. And we should have the ability to transparently and securely use different stores with any kind of analytics."

"Sometimes you can't move data or aggregate all the data that may be useful to your particular agency," Johnson added. "There are several examples where an agency said 'I'd like this from this other agency, but for a variety of reasons, sometimes laws, privacy restrictions like HIPAA, they can't give it to us.' Many agencies would like to have ways of reaching across different datasets and federating queries." By developing a properly designed Big Data analytics architecture, agencies can securely provide an analyst access to the right information, at the right time.

Another challenge agencies face? Keeping costs low by leveraging open source products to power big data programs, but avoiding expensive costs related to integration and consulting. The key for agencies is to build a big data infrastructure that is balanced between open source and proprietary solutions to simplify setup and use.

"There's this idea that open source can do everything," said Johnson. "But there are better ways to set up, maintain and secure Big Data architectures. The great thing about open source is you can use it to drive down your costs, but you should only use it where it makes sense and determining where that is what we help our government partners with."

One example of where Big Data can make government more effective for citizens comes from a project from the National Cancer Institute. They were looking to match 17,000 genes known to be related to certain cancers with the nearly 20 million research articles in the PubMed medical library. This data would then be used to create customized treatment plans, based on an individual's genome across a population of 900 Million citizens.

"Because there's 20 million text abstracts, you're searching basically unstructured data," explained Johnson. "You do have some structured datasets in the gene types, the known cancer associations, but matching all that different data up was something they simply couldn't solve, until we brought in this idea of using big data. And a key part of it was open source."

Johnson said a critical component was making sure they were using the right open source tools with other product to make the process work. "Rather than try to string together a whole cluster and download software and install it, we could bring in our tools that made it very quick to set up. We started the process on a Friday, and on Monday morning the National Cancer Institute had their answer," said Johnson.

The National Cancer Institute is one of many examples of how Oracle is helping agencies capitalize on their data. With Oracle, public sector leaders can obtain a complete, open and secure suite of big data technologies, servers, and storage solutions engineered to work together, optimizing every aspect of government operations.

what IT is POWERING your big data initiative?

LEARNING OBJECTIVE

This section will give you an overview of the big data technology landscape and help you understand what IT you need. You'll also gain a better understanding of available IT solutions.

We've talked about what big data is, how to deploy it at your agency, and the people needed to run a program, but what kind of IT is needed to support your initiative? What does the current big data market look like? We'll discuss options and give you some guidance on what kinds of IT investments you might be looking at.

Across government, there is a dire need to invest and upgrade IT infrastructure. To fully leverage big data, your agency must have the right infrastructure to support these initiatives. This means the ability to quickly process volumes of data and run reports customized to fit your agency's unique mission needs.

Here, we provide a quick overview of some of the IT components needed to power big data. Although this is not a fully inclusive list, the following solutions are a great start to build your big data IT ecosystem.

IT Component: Cloud Computing

Cloud computing is very likely already on your IT roadmap. Like big data, it can be defined in many ways. We've settled on defining it in the following way: an IT delivery model that enables on-demand access to resources, whether it is access to servers, computing power, applications, data or software, which can then be quickly acquired by employees across laptops, desktops or smartphones. Within that definition, there are a variety of cloud service and deployment models, which are highlighted throughout this section. With cloud, you will be able to quickly deliver information and IT services to power your big data initiative.

3 Common Cloud Service Models

Below, we identify three of the most common cloud service models.

SOFTWARE AS A SERVICE (SAAS)

Software as a Service (SaaS) is a cloud service model in which an agency accesses software on demand, from a third party vendor. The agency does not buy the software but is provided multiple licenses to access information. Examples of SaaS include:

- Blogging, social networking and online communications platforms
- Online email services

PLATFORM AS A SERVICE (PAAS)

Platform as a Service (PaaS) is a cloud delivery model where a vendor provides an online development platform for an agency. Developers leverage the vendors' computing environments and can test, create and ultimately host new applications. Examples of PaaS:

- Application design, development and deployment
- Team collaboration solutions

INFRASTRUCTURE AS A SERVICE (IAAS)

Infrastructure as a Service (IaaS) is a cloud delivery model where a vendor provides the hardware and software, and a government agency can build a customized computing environment. This delivery model can provide government agencies with access to advanced computing power, storage, memory, bandwidth and software applications — all available on demand. Examples of IaaS:

- Operating systems, servers and storage capabilities
- Networking components and software bundles

Cloud Deployment Models

Generally, there are four kinds of deployment models, which we describe below.

THE PUBLIC CLOUD

A public cloud is a cloud deployment that makes information available for the public.

THE PRIVATE CLOUD

A private cloud is a cloud deployment that is used exclusively for internal applications within an agency, but multiple business units may be granted access to share information and manage data.

THE HYBRID CLOUD

The hybrid cloud consists of two or more deployment models. For instance, a hybrid cloud will contain both a public and private cloud and has the ability to easily segment data and transfer data between clouds as necessary.

THE COMMUNITY CLOUD

A community cloud is a cloud deployment model that provides access to multiple organizations that have a similar interest in collaboration. You may also hear this kind of cloud referenced as a "government only" cloud model.

software AG

Government Solutions

Take a “Special Forces” Approach to Enterprise IT

Helping Agencies Prove IT First and Prove IT Fast

Software AG Government Solutions is a leading software solutions company, delivering massive-scale, complex and real-time solutions for:

- Business Process Management
- Integration
- Analytics & Visualization
- Application Optimization

Put our team to the test.
Learn how our "special forces" approach can get fast results for your most complex integration and process challenges. Visit: www.SoftwareAGgov.com

GET THERE FASTER

Unlock New Insights by Breaking Down Old and New Data Silos

AN INTERVIEW WITH CHRIS STEEL, CHIEF SOLUTIONS ARCHITECT, SOFTWARE AG

With big data and analytics programs, agencies can reveal new insights and improve their decision-making process in ways they were unable to before. But to unlock these new insights from information, government agencies must first understand how to fully leverage all their high value and authoritative data. This means that in order to truly create a data-driven agency, the way government stores, accesses and processes data must change.

"We see a lot of data within agencies stored across different information silos," said Christopher Steel, Chief Solutions Architect at Software AG. "One of the biggest hurdles to getting all of the data together is the fact that it's often stored in different formats."

Steel explained that there are two common data categories that agencies collect: static and streaming. Static data refers to information that is stored in a traditional database or data warehouse or Hadoop and analysis is run on the information. Streaming data refers to real-time data, with information coming from sensors, devices or social media. In order to maximize the most benefits, agencies must be able to run analysis on both static and streaming data.

"Organizations should be able to take data that's in one form and convert it to a form that's needed, and share that data in real-time," said Steel. "This empowers leaders to make decisions they previously weren't able to make. This entire process can be done in real-time, whereas before it was either not possible, because there was just no mechanism to get the data into the format needed, or it would take a relatively long amount of time."

Government agencies must have the ability to extract value from static data. This data might be stored in a system like Hadoop. But to get a full understanding of an issue, static data should merge with real-time data to for a complete analysis. "If you have a Hadoop solution where you're running analysis on an hourly, daily, weekly basis, what Software AG can bring to the table is the ability to augment those analysis with real-time data. We can color that data with contextual information from real-time feeds," said Steel.

Even when bringing information together, there is still the need to put real-time data into the proper context to measure against historical trends. For instance, transactional systems host troves of historical data, but now are also producing volumes of real-time data. When real-time data is measured against historical trends, organizations can quickly spot abnormalities, and may be able to thwart waste, fraud and abuse –in some cases, before it even happens.

Another example of why it's important to blend historical and real-time data comes from when a cyberattack occurs. When this happens, agencies are flooded with data from log files. Often, much of this data is irrelevant, so an agency needs an effective way to sift through the data to quickly identify the abnormality in the log file and provide insights into the breach.

"With real-time big data analytics, agencies can pick out the needles in the haystack as data is streaming by, and they don't have to worry about taking al of the irrelevant data, putting it into a store, running an analysis, waiting, and then redoing that over and over again," said Steel.

"Instead, in real-time, the agency can have the ability to pull information out and make a decision while they still have a window of relevance, which allows them to act on the data. Once the agency sees that the attacker has breached their systems, they can go in and take proactive action to stop the hacker from stealing or corrupting any of data."

Added Steel, "In the past, agencies were behind the curve. They might be able to eventually figure out that a breach had occurred from the the information, but by that time, their data was already stolen or corrupted."

But by exploring both the historical and streaming data, agencies can gain the ability to analyze data to optimize operations, mitigate risks, and make decisions in real-time.

"For over a decade, Software AG has been working with agencies to help break down their information silos," said Steel. "We do that through a variety of different products that allow public sector agencies to integrate, transform and make data more accessible and faster to help reduce costs."

The BIG DATA IT Glossary

Before you have an understanding of the terms associated with a big data solution, let's review what steps should you take to deploy a big data infrastructure? We highlight three essential steps below.

STEP 1: IDENTIFY YOUR STORAGE NEEDS

One of the first steps will be scoping out what you need in terms of storage. There are various options out there, but in terms of needs, you must be a smart consumer. Be sure that your purchase is scalable, so if you need more storage, you can quickly access more and then reduce it as demand decreases.

STEP 2: PICK THE RIGHT STORAGE MEDIUM

Once your needs are assessed, you then have to map your needs to the right storage medium. This could be cloud, servers or flash storage.

STEP 3: MAKE SURE YOUR INFRASTRUCTURE IS DATA CENTRIC

Your big data infrastructure should make it easy for anyone to access data. So the infrastructure you select should create an efficient and flexible data platform. This means that it is easy to search, query and store data at your agency.

Looking for more information? Be sure to check out Worksheet 4: Your Big Data Infrastructure Checklist. But for now, here are some questions you can ask:

1. What is our current data architecture? How can it be improved?
2. Are we laying a foundation for future data initiatives? How do we scale as needed?
3. What kind of data will we need to extract in the future? How can we think differently about the data we already have?
4. What kind of data ecosystem is needed to share data across our agency?
5. How do we secure the privacy of data?
6. What can we do to reduce insider threats?
7. What feedback mechanisms do we have?
8. Are we delivering on the needs of our users?

HADOOP

Hadoop, formally called Apache Hadoop, is an Apache Software Foundation project and open source software platform for scalable, distributed computing. Hadoop can provide fast and reliable analysis of both structured data and unstructured data. Given its capabilities to handle large data sets, it's often associated with the phrase big data.

The Apache Hadoop software library is essentially a framework that allows for the distributed processing of large data sets across clusters of computers using a simple programming model. Hadoop can scale up from single servers to thousands of machines, each offering local computation and storage.

Definition from:
http://www.webopedia.com/TERM/H/hadoop.html

HADOOP MAPREDUCE

Hadoop MapReduce (Hadoop Map/Reduce) is a software framework for distributed processing of large data sets on compute clusters of commodity hardware. It is a subproject of the Apache Hadoop project. The framework takes care of scheduling tasks, monitoring them and re-executing any failed tasks.

According to The Apache Software Foundation, the primary objective of MapReduce is to split the input data set into independent chunks that are processed in a completely parallel manner. The Hadoop MapReduce framework sorts the outputs of the maps, which are then input to the reduce tasks. Typically, both the input and the output of the job are stored in a file system.

Definition from:
http://www.webopedia.com/TERM/H/hadoop_mapreduce.html

MASSIVELY PARALLEL PROCESSING

Massively parallel processing (MPP) is a form of collaborative processing of the same program by two or more processors. Each processor handles different threads of the program, and each processor itself has its own operating system and dedicated memory. A messaging interface is required to allow the different processors involved in the MPP to arrange thread handling. Sometimes, an application may be handled by thousands of processors working collaboratively on the application.

Definition from:
http://www.techopedia.com/definition/2786/massively-parallel-processing-mpp

There are a lot of key terms used when discussing big data. Here, we identify some of the most important words and phrases you need to understand when making your IT investment.

NOSQL

A NoSQL database provides a mechanism for storage and retrieval of data that is modeled in means other than the tabular relations used in relational databases. Motivations for this approach include simplicity of design, horizontal scaling and finer control over availability. The data structures used by NoSQL databases (e.g. key value, graph or document) differ from those used in relational databases, making some operations faster in NoSQL and some faster in relational databases. The particular suitability of a given NoSQL database depends on the problem it must solve.

Definition from:
http://en.wikipedia.org/wiki/NoSQL

APACHE HBASE

Apache HBase (HBase) is the Hadoop database. It is a distributed, scalable, big data store. HBase is a subproject of the Apache Hadoop project and is used to provide real time read and write access to your big data. According to The Apache Software Foundation, the primary objective of Apache HBase is the hosting of very large tables (billions of rows X millions of columns) atop clusters of commodity hardware.

Definition from:
http://www.webopedia.com/TERM/A/apache_hbase.html

APACHE HIVE

Apache Hive (Hive) is a data warehouse system for the open source Apache Hadoop project. Hive features an SQL-like HiveQL language that facilitates data analysis and summarization for large data sets stored in Hadoop-compatible file systems.

Hive originated as a Facebook initiative before becoming a subproject of Hadoop. Hive is currently an open source volunteer top-level project under the Apache Software Foundation.

Definition from:
http://www.webopedia.com/TERM/H/hive.html

APACHE PIG

Apache Pig is a high-level procedural language platform developed to simplify querying large data sets in Apache Hadoop and MapReduce. Apache Pig features a "Pig Latin" language layer that enables SQL-like queries to be performed on distributed data sets within Hadoop applications.

Pig originated as a Yahoo Research initiative for creating and executing MapReduce jobs on very large data sets. In 2007, Pig became an open source project of the Apache Software Foundation.

Definition from:
http://www.webopedia.com/TERM/A/apache_pig.html

big data in the WINDY CITY

The city of Chicago has long been a leader in government innovation. The city won funding from the Bloomberg Philanthropies' Mayors Challenge to build a predictive policing platform. They've created an open data platform, and now they've embarked on a new program called "The Array of Things."

This program seeks to build a network of interactive sensors to collect real-time data about the city — everything from environmental concerns to infrastructure and then making the data open and accessible through their open data platform.

"The data [from 'The Array of Things'] will be in contrast to most of the data we already have, which is administrative data (related to HR), finance, 911 and 311 calls. Data being collected through 'The Array of Things' is very different than data collected through our administrative process," said Tom Schenk, Chief Data Officer, city of Chicago. "What we are able to do [with big data] is take a lot of work that takes place in the city and be able to sort it to be more efficient."

One example of big data success is how the city has been able to control its rat population. "Using data, we can forecast where rodents are going to be and route crews to that location," said Schenk. Citizens can also report locations where they saw a rat, and city officials can go to remove the infestation, or clear out the area to make it less prone to rats.

Schenk identified a few challenges that organizations often face when starting their big data programs. One is gathering data from multiple sources. Another initial obstacle for big data is finding the right program and scoping the research project.

"In other research venues I have worked in, I have been able to reference decades of research, to the point where you can take research that has been done and apply it to your particular city," he said. "But in municipal research, the problem that exists is the literature does not align with the actual problems of a city."

In order to be sure they are working on the right projects that would benefit employees the most, Schenk and his team often ask their colleagues the question: How can I use data to help you?

"The person I am usually asking is not a data person. He or she is not a researcher. So it takes a little work to get the question out in the right way," said Schenk. Because of this, Schenk recommended that when getting started with big data, you must first focus on building relationships. This will allow teams to understand what problems to tackle and how analytics can do it.

"It takes a lot of conversation. As soon as you start to get into the details of the complexity of projects, you start to understand workflows, so I would focus on communicating with those business owners and trying to understand their problems," said Schenk.

PREPARING for the big data journey

LEARNING OBJECTIVE

Now that you've worked your way through the guide, we want to provide you with some clear action items and templates to help you create your big data strategy.

1. Creating your big data statement of purpose

Our playbook has talked about the importance of creating a big data statement of purpose. Follow this worksheet to start building your own.

2. Overcoming common big data challenges

Are you worried that you're going to run into challenges even before deploying your big data program? Take a look at this quick chart to help walk you through some common obstacles. Use this as a way to have a discussion with your team, and tailor the scenario to your particular agency. #3:

3. Building your big data team workforce planner

Like we mentioned in the report, you might not be able to hire new staff, but you can assess your workforce for gaps. Use these charts as a beginning to assess where skills gaps might be, and then you can consider building out training opportunities or leverage organizational talent in new ways.

4. Your big data infrastructure checklist

We've included some very quick ideas on what your big data infrastructure should be able to do.

5. Building your 60-day action plan

Now that your statement of purpose is complete, you'll have to focus on executing the vision. So in this worksheet, we're going to build a 60-Day Action Plan.

Notes

WORKSHEET 1

CREATING YOUR BIG DATA STATEMENT OF PURPOSE

STEP 1

Start by answering these ten questions, then build out your statement of purpose.
We'll walk you through how to do that below.

1. What problem are we trying to solve? How will data help?

2. How will big data work to meet your mission needs?

3. What outcomes do you want to achieve?

4. What kinds of data do we need access to?

5. Who are the main stakeholders and how do we engage them?

6. How are we going to track, assess and monitor progress?

7. Can we start small by piloting a few programs? What can we learn from starting small and building out?

8. Do we have the right workforce in place?

9. Have we received buy-in from leadership and across teams?

10. Are we delivering on the needs of our users? How do we know?

STEP 2

Now that you've answered these questions, you're ready to build your statement of purpose.

Executive Overview

A brief overview and high level description of what the project will be.

Impact and Benefits

A discussion on what the expected deliverables will be, and your expected outcomes.

Key Milestones & Timelines

A discussion of when it will be achieved and an anticipated timeline.

Notes

WORKSHEET 2

OVERCOMING COMMON BIG DATA CHALLENGES

Are you worried that you're going to run into challenges even before deploying your big data program? Take a look at this quick chart to help walk you through some common obstacles. Use this as a way to have a discussion with your team, and tailor the scenario to your particular agency.

CHALLENGES & SOLUTIONS	ACTIONS TAKEN	CHALLENGES & SOLUTIONS	ACTIONS TAKEN
Lack of Leadership Support Solutions ☐ Show the Business Value ☐ Show Use Cases or Examples ☐ Create communities of practice		**We've Hit A Cultural Roadblock** Solutions ☐ Build a Culture of Trust ☐ Allow for Risk, Embrace Failures ☐ Focus on Education and Training	
Slow Acquisition Process Solutions ☐ Consider your entire big data infrastructure design ☐ Plan for future when purchasing ☐ Explore shared services models or leverage existing IT		**Your Agency:** Solutions ☐ ☐ ☐	
Lack of Clarity on Data Needs Solutions ☐ Conduct a data audit ☐ Prioritize Data ☐ Don't Reinvent the Wheel		**Your Agency:** Solutions ☐ ☐ ☐	
Migrating Our Data and Applications Efficiently Solutions ☐ Define personas to manage data ☐ Governance is essential ☐ Collaborate across teams to avoid downtime		**Your Agency:** Solutions ☐ ☐ ☐	

Notes

WORKSHEET 3A

BUILDING & ASSESSING YOUR BIG DATA TEAM

HOW TO USE THIS WORKSHEET:

This sheet should be used as a quick checklist to help focus your big data journey. With your team, ask each question in the "Discussion Starters" column. Once you have a thorough discussion, check the box and advance to the next question. Along the way, be sure to assign a note taker. Use your notes to help build your roadmap, and develop more thorough strategies. Your answers will also be important as you build your 60 day action plan.

PLANNING ACTIVITES & DISCUSSION STARTERS

1. Review Your Organization's Big Data Statement of Purpose
   - ☐ Do we have an in-depth knowledge of project requirements?
   - ☐ Have we considered resource needs, like time & labor?
   - ☐ Do we know staffing requirements for this project?

2. Assess Existing Workforce
   - ☐ Have we defined how many current full-time employees (FTE) we need?
   - ☐ Do we know how many current contractors do we use?
   - ☐ Can we estimate the number of retirees in the next 5 years?

3. Create a Model of Future Workforce
   - ☐ Do we know skill sets needed for future?
   - ☐ Have we identified the specific skills to meet project goals?

4. Identify Necessary Skills
   - ☐ Have we addressed our data skills gap?
   - ☐ Do we know if a skills gap exists?

5. Hire or Provide Training
   - ☐ Can we provide training to our current staff?
   - ☐ Can we conduct virtual trainings?
   - ☐ Do we have a budget to hire?

6. Incentives and retain talent
   - ☐ Do we have a strategy to retain top talent?
   - ☐ Do we have any mentorship programs?

DISCUSSION STARTERS	COMPLETED
Do we have an in-depth knowledge of project requirements?	☐
Have we considered resource needs, like time & labor?	☐
Do we know staffing requirements for this project?	☐
Have we defined how many current full-time employees (FTE) we need?	☐
Do we know how many current contractors do we use?	☐
Can we estimate how many retirees there will be in the next 5 years?	☐
Do we know skill sets needed for future?	☐
Have we identified the skills that are specific to meet project goals?	☐
Have we addressed our data skills gap?	☐
Do we know if a skills gap exists?	☐
Can we provide training to our current staff?	☐
Can we conduct virtual trainings?	☐
Do we have a budget to hire?	☐
Do we have a strategy to retain top talent?	☐
Do we have any mentorship programs?	☐

Notes

WORKSHEET 3B

BUILDING & ASSESSING YOUR BIG DATA TEAM

DISCUSSION STARTERS

	Big Data Skill Sets	In-House?	Outsourced?	Missing Skill?	Employee(s)	Actions
KNOWLEDGE WORKER	Expertise in market trends on data and analytics.	☐	☐	☐		
	Articulates data and findings across teams and to senior management.	☐	☐	☐		
	Provides insights on potential impacts from data-driven policies.	☐	☐	☐		
	Communicates to external and internal stakeholders about program success.	☐	☐	☐		
	Promotes and advertise the big data program across your team.	☐	☐	☐		
	Works across teams as a liaison, speak IT and business language.	☐	☐	☐		
	Gathers feedback on citizen and user needs.	☐	☐	☐		
	Assists citizens as they develop applications and tools from your open data programs.	☐	☐	☐		
	Encourages your agency to adopt citizen science initiatives for additional data collection.	☐	☐	☐		
	Documents, track and manage big data programs.	☐	☐	☐		
MANAGEMENT ROLE	Operates as a strong team builder, exceptional interpersonal skills and can speak both IT and business language effectively.	☐	☐	☐		
	Provides strategic direction for data applications across the agency.	☐	☐	☐		
	Possess a thorough understanding of business strategy and the link to organizational data.	☐	☐	☐		
	Works across teams to learn about initiatives and impact on organization.	☐	☐	☐		
	Understanding of agency mission and importance of data.	☐	☐	☐		
	Serves as project lead, making sure projects stay on time and on budget.	☐	☐	☐		
	Consults and provides advisement on technology needs.	☐	☐	☐		
TECHNICAL SKILLS	Creates, manage and operate databases for big data programs.	☐	☐	☐		
	Expertise in big data technologies (Hadoop, for example).	☐	☐	☐		
	Understands cloud computing technology	☐	☐	☐		
	Knowledge of programming languages and graphic design techniques.	☐	☐	☐		
	Extracts key data and working across teams to understand visualizations.	☐	☐	☐		
	Can deploy world-class IT and infrastructure to programs	☐	☐	☐		
	Develops cutting edge technology to push big data programs further.	☐	☐	☐		
	Knowledge to query databases and conduct statistical analysis.	☐	☐	☐		

EMPLOYEE INFO

Knowledge worker: Think about your knowledge workers, mix of project management and work done by a subject matter expert

Management Position: A bit more senior role for the organization, can communicate needs up the chain, purchasing power.

Technical Skills: Will require a very particular skillset, like a data scientist.

WHY ARE THESE IMPORTANT?

So why are these important to consider? These factors:

1. Identify potential employees to fill your skills gap

2. Identify potential trainings and educational opportunities

3. Define actions your agency can make to develop needed skills

BIG DATA INFRASTRUCTURE CHECKLIST

Below are some quick ideas on what your big data infrastructure should be able to do. * Start this **AFTER** you've filled out the rest of the worksheet.

Can Your Infrastructure Do the Following?	Y/N
Provide accessibility to information	☐
Integrate with existing systems	☐
Leverage existing IT systems	☐
Conduct advanced and operational analytics	☐
Collect, store and manage any kind of data type or file	☐
Easily search for information	☐
Secure the integrity of your data	☐
Scale to meet demands	☐
Create an easy to use platform and interface	☐
Run advanced analytics for business units	☐

Notes

WORKSHEET 4

YOUR 60 DAY ACTION PLAN PREP SHEET

Let's do some quick prep work in order to build your plan. Now that your statement of purpose is complete, you'll have to focus on executing the vision. So we're going to build a 60 day action plan.

ACTION PLAN PREP QUESTIONS

1. Are we risking scope creep? Are we staying true to our stated objectives?	Y/N
We've defined our state of purpose	☐
We've established meetings and discussed projects	☐
We've assigned a project manager	☐

2. Have you audited your data?	Y/N
We've collaborated across teams and created a data inventory	☐
We've prioritized our high value data	☐
We have buy-in from business units to share data	☐

3. Have we tested and re-tested data migrations?	Y/N
We have tested migrations to spot any bugs	☐
We have notified teams about when downtime may occur	☐
We have data protections in place when moving information	☐

4. Have we collaborated with related teams to build a desired solution?	Y/N
We have clearly defined our end users	☐
We are building a solution that is desired	☐
We've gained buy-in from teams for the big data program	☐

5. Have we discussed our governance policy?	Y/N
We have reviewed our data governance policies	☐
We have trained staff on acceptable data usage	☐
We have set meetings to review our governance policy	☐

DAYS 1-15

Research Discovery & Deployment		Actions Taken
Define the core problem you want to achieve with big data	☐	
Identify how big data solves your needs	☐	
Select a core problem to solve and use as big data pilot	☐	
Identify your core users for that problem	☐	
Identify your top five challenges to overcome	☐	
Develop your statement of purpose	☐	
Have project manager defined to kick off program	☐	

DAYS 16-30

Audit, Prioritize & Communicate		Actions Taken
Define the core problem you want to achieve with big data	☐	
Identify how big data solves your needs	☐	
Select a core problem to solve and use as big data pilot	☐	
Identify your core users for that problem	☐	
Identify your top five challenges to overcome	☐	

DAYS 31-45

Assess, Prepare & Identify Needs		Actions Taken
Conduct a workforce assessment for your program	☐	
Identify skills gaps for agency	☐	
Identify strategies to find training	☐	
Assess your infrastructure needs	☐	

DAYS 46-60

Deploy, Finalize & Iterate		Actions Taken
Begin preliminary strategies to deploy program	☐	
Work with team to develop infrastructure	☐	
Find a quick win to continue support	☐	
Finalize your data governance policy	☐	
Collaborate across team	☐	
Conclude project and identify additional programs	☐	
Debrief for lessons learned and ways to improve	☐	

About GovLoop

GovLoop's mission is to "connect government to improve government." We aim to inspire public-sector professionals by serving as the knowledge network for government. GovLoop connects more than 150,000 members, fostering cross-government collaboration, solving common problems and advancing government careers. GovLoop is headquartered in Washington, D.C., with a team of dedicated professionals who share a commitment to connect and improve government.

For more information about this report, please reach out to info@govloop.com.

GovLoop
1101 15th St NW, Suite 900
Washington, DC 20005
Phone: (202) 407-7421
Fax: (202) 407-7501
www.govloop.com
Twitter: @GovLoop

Acknowledgments

Thank you to Arrow Electronics, Cisco Systems, Cloudera, NetApp, Oracle, and Software AG for their support of this valuable resource for public-sector professionals.

AUTHOR:

Patrick Fiorenza, GovLoop's Senior Research Analyst

DESIGNERS:

Jeff Ribeira, GovLoop's Senior Interactive Designer
Tommy Bowen, GovLoop's Junior Designer
Kaitlyn Baker, GovLoop's Design Fellow

EDITOR:

Catherine Andrews, GovLoop's Director of Content

1101 15th St NW, Suite 900
Washington, DC 20005

Phone: (202) 407-7421 | Fax: (202) 407-7501

www.govloop.com
Twitter: @GovLoop