insideBIGDATA

*InsideBIGDATA Guide to*

# Data Analytics in Government

*Written by Daniel D. Gutierrez, Managing Editor, insideBIGDATA*

BROUGHT TO YOU BY DELL EMC | intel

## Introduction

*This guide provides an in-depth overview of the use of data analytics technology in the public sector. Focus is given to how data analytics is being used in the government setting with a number of high-profile use case examples, how the Internet-of-Things is taking a firm hold in helping government agencies collect and find insights in a broadening number of data sources, how government sponsored healthcare and life sciences are expanding, as well as how cybersecurity and data analytics are helping to secure government applications.*

## Data Analytics for Government: An Overview

Data is a critical asset for government agencies. All levels of government collect an increasing amount of data every day. As these agencies strive for a meaningful digital transformation, it's important not only to collect and store large data sets, but also to use that data when making mission-critical decisions. This growing use of so-called "big data" builds mounting pressure on government to use data analytics to turn all data into actionable information. Efficient use of data is the missing link between good governance and capacity building, where data insights can be gleaned to improve service delivery. This technology guide will help government thought leaders in how best to use data analytics to manage and derive value from an increased dependence on data.

A number of prevailing concerns include:

- How are these data is being converted into beneficial insights?
- What happens when you have too much data?
- How do you make sense of it when the data volume is on a continued upward trajectory?
- How can you keep up with the volume of data?

Government also must look at the type and source of data being collected, stored, analyzed, and consumed, i.e. structured versus unstructured data. Structured data is the data that has been modeled and normalized to fit into a relational model, such as traditional row/column databases. Unstructured data is information that either doesn't have a predefined data model or doesn't fit well into relational tables — examples are social media, plain text, log files, video, audio, and network-type data sets. Big data for data analytics is a compilation of both structured and unstructured data.

Recognizing the importance of data analytics, the U.S. Government now has a new official role of *Chief Data Scientist*. Pledging to put all government records into the public domain, it is clear that U.S. Government officials are far from ignorant of the importance of the data revolution.

### Contents

**KEY TAKEAWAYS**

- Data analytics for government is a rapidly evolving field
- Government agencies have an appetite for embracing big data technologies
- Government leaders need to focus on delivering value

Certain uses are already being found for data in government-related fields, such as healthcare, cybersecurity, and education, which can potentially have huge positive impact. These include:

- The CIA helped to fund Palantir Technologies, which produces analytical software designed to stamp out terrorism and crack down on cyber fraud by identifying transactions that follow patterns commonly displayed during fraudulent activity.
- American law enforcement agencies (at federal, state and county levels) have access to sophisticated ALPR (automated license plate recognition) software that alerts them to anyone in the vicinity with an outstanding warrant. And predictive technologies are used by several police departments to predict flashpoints where crimes may occur, as well as link particular crimes to particular repeat offenders.
- The U.S. Department of Transportation also uses license plate recognition, as well as cameras, to monitor the flow of people as they travel by plane, train and automobile, generating insights as to where infrastructure investment is necessary, as well as predictions about how we are likely to change the way we travel in the future.
- In education settings around the globe, thanks to the huge increase in the amount of learning activity carried out online (both in traditional school environments and through distance learning), massive amounts of data about the way we study and learn is becoming available.
- The U.S. Department of Agriculture supports the agricultural economy through research and development of new big data technologies. One recent breakthrough is increasing the yield of dairy herds by identifying bulls most likely to breed high-yielding cows based on genetic records.
- Government agencies involved in healthcare, such as the Centers for Disease Control (CDC), track the spread of illness using social media, while the National Institutes of Health (NIH) launched an initiative called *Big Data to Knowledge* (BD2K) in 2012 to encourage innovation based on data-derived insights. One project funded by the government even aims to spot the early signs of suicidal thinking among war veterans, based on their social media behavior.

Privacy concerns are often seen as the biggest challenge presented by the rise of data and analytics, and the U.S. Government has identified several potential areas that could be infringed on through inappropriate use of data, including rights to secrecy of personal information and rights to remain anonymous while exercising free speech. Indeed, big data use in government certainly presents big challenges — officials and politicians have a fine line to tread if they do not want to come across as attempting to implement a real-life version of Orwell's Big Brother.

## Emerging Technologies for the Public Sector

A new report by Accenture, *Emerging Technologies in Public Service*, examines the adoption of emerging technologies across government agencies, with the most interaction with citizens or the greatest responsibility for citizen-facing services: health and social services, policing and justice, revenue, border services, pension and social security, and administration. A few of the findings of this nine-country survey of nearly 800 IT leaders include:

- More than two-thirds of government agencies are evaluating the potential of emerging technologies including big data, but only one-fourth have moved beyond the pilot stage
- Those that have embraced technologies like IoT and machine learning all report that the most common new IT used is advanced analytics with predictive modeling
- Nearly half of those agencies say their primary objective in harnessing the power of advanced analytics is to improve and support employee work
- More than three-fourths indicate that implementing machine learning methods are either underway or complete

## Preparing for Big Data

What can government agencies do to prepare for and embrace big data and data analytics? First of all, government agencies should try to get ahead of their data deluge. Strategy, planning and governance are critical to this process. Second, they must develop and review the life cycle for data for their agencies. The life cycle can be categorized into the following phases:

### Data Life Cycle Phases

Retail generates a flood of complex structured and unstructured data. There is a vast number of sources of this data, but for a short-list we can consider the following:

- **Ingest:** the collection of data from a diverse set of sources
- **Store:** the repository for the collected data — the right kind of data needs to be stored in the correct repository
- **Analyze:** the analytics of the data in the repositories
- **Consume:** the reporting and business intelligence for decision-making

When the big data life cycle is well understood, government thought leaders need to plan and identify the following:

- **Find technology enablers –** seek out and evaluate new infrastructure and new software applications, and institute pilot programs/ proof-of-concept projects
- **Adopt an ecosystem approach –** big data analytics is an evolving space and there will be new technology options to review and select new solutions
- **Adopt a use case-based approach –** data's value depends on the insight of the domain; look for use case-specific projects, e.g. network-centric data analytics or cybersecurity insights
- **Invest in data-centric skill sets –** insights in large data sets tend to be as good as the domain knowledge of the data, so skills for data scientists and data analysts need to be developed and nurtured

The Japanese government plans to develop a system to help disaster victims by utilizing big data gathered from sources such as internet postings, and global positioning system (GPS) data from smartphones and car navigation devices.

In setting the stage for successful big data deployments, government agencies also need a good sense for data lineage — source of data, change log of data, and trustworthiness of data to limit any sort of garbage-in-garbage out (GIGO) possibilities.

## Innovative Use Case Examples

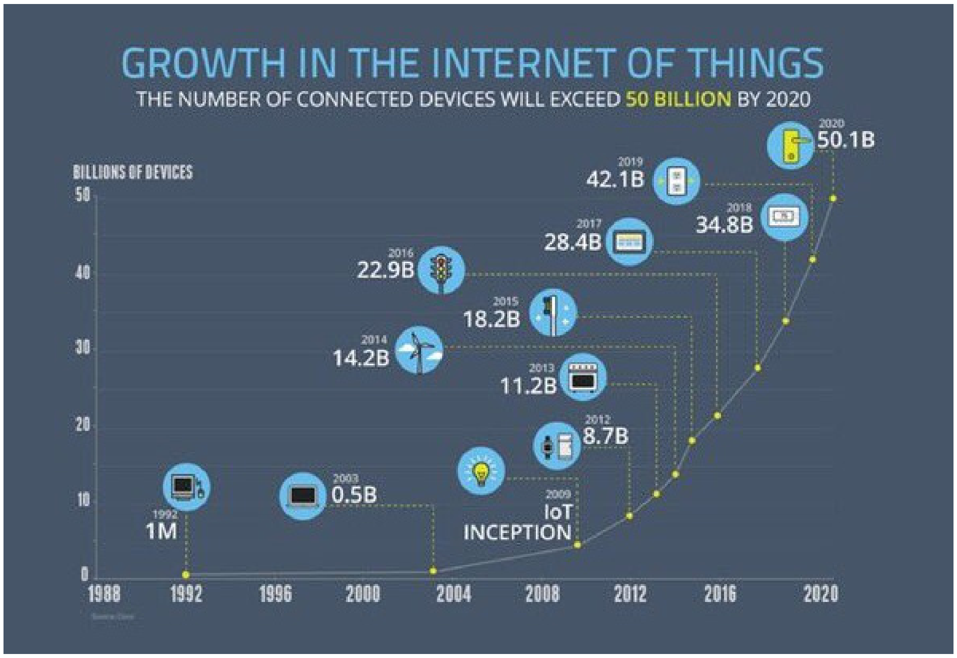
**There are a number of high-profile big data use case examples in a government setting:**

- Pakistan's National Database & Registration Authority (NADRA), one of the world's largest multi-biometric citizen databases, serves as an example of harnessing the power of government data. NADRA is an independent and autonomous agency under Ministry of Interior and Narcotics Control, Government of Pakistan that regulates government databases and statistically manages the sensitive registration database of all the national citizens of Pakistan.
- The Japanese government plans to develop a system to help disaster victims by utilizing big data gathered from sources such as internet postings, and global positioning system (GPS) data from smartphones and car navigation devices. The system will enable administrative authorities to immediately ascertain the movements of victims just after a disaster occurs. The government will gather and analyze information, including that on isolated local communities and overcrowded shelters, to make the initial response after a disaster, such as search and rescue operations and the delivery of goods, more efficient.

- From Washington, DC to cities, states, and countries all around the world, "open government" is revolutionizing the way citizens interact with government leaders. It connects like-minded citizens with each other, with government agencies, and with many other types of organizations. To support open government initiatives and uphold the values of transparency, participation, and collaboration in the US, federal agencies now make their data open. This means making large data stores publicly accessible in a format that can be shared. Open data from the government gives citizens the information they need to hold government leaders accountable. Open data fosters collaboration between government leaders and citizens, and encourages cooperation internally among government entities. The results can be tremendously better decisions that have the potential to drastically change lives. One fertile source of open data from the Federal Government, as well as APIs from a variety of federal agencies and other resources, is data.gov.
- A new UKAuthority report "Digitising Policing," indicates that advances of IT solutions like data analytics are changing the world and policing is changing with it. It is clear that police forces are rapidly adopting big data technology. It promises significant improvements in efficiency — the management of evidence can be improved, and the provision of more information — can support better decision making at strategic and operational level, and among officers on the beat. There are also opportunities to use data analytics to better understand factors influencing the demands on the police.
- Traffic and congestion control in London has been using automatic plate number recognition along with multiple video cameras across the city for years meaning that during the planning and delivery of Olympics 2012, Transport for London (TfL) employed a number of big data tools to make sure that during the Olympic games public (and private transport) kept people moving quickly and safely to and from the games and generally around London for their normal business.

## Government Applications of IoT

The *Internet of Things* (IoT) is a term used to describe the set of physical objects embedded with sensors and connected to the Internet. IoT offers numerous opportunities for the Federal Government to cut costs and improve citizen services. The data visualization "Growth in the Internet of Things" (below) shows the tremendous growth in the number of connected devices across all sectors.



Source: NCTA – The Internet & Television Association

## Smart Cities

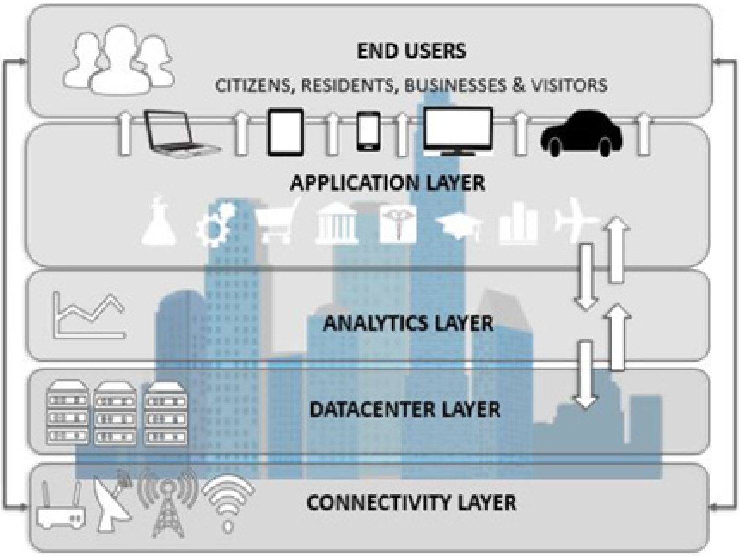
IDC defines a Smart City (also known as digital city) as being a *"finite entity (district, town, city, county, municipal, and/or metropolitan area) with its own governing authority that is more local than national. This entity is built on an information and communication technology (ICT) foundation layer that allows for efficient city management, economic development, sustainability, innovation, and citizen engagement."*

Data is one of the most critical elements that will underpin the success of a city's transformation into a smart city. To be deemed successful, a city should be able to harness data from existing government systems, online and mobile applications, third-party applications, and, most importantly, from citizens — the ultimate beneficiaries of smart cities. The data that is gathered can be analyzed and used to make informed, automated decisions that can improve the life of citizens and better manage local resources.

Smart cities are communities that harness technology to transform physical systems and services in a way that enhances the life of its residents and business while making government more efficient. It is more than a mere automation of processes; it also links disparate systems and networks to gather and analyze data that is then used to transform whole systems. While the idea behind smart cities has been around for years, it has acquired a new urgency as more people move into urban centers. Sustainability is becoming an important imperative and technologies have advanced to a point where there can be real-time and meaningful interaction between cities, residents, and businesses.

According to the U.S. Smart Cities Council, another perspective of a smart city is a "system of systems — water, power, transportation, emergency response, and others — with each one affecting all the others." In recent years, the ability to merge multiple data streams and analyze them for critical insights has been refined. Those insights are what enhance livability, workability, and sustainability in a city. The overarching goal is to make cities more efficient, and ensure resources are found where they're needed.

### Building Blocks of a Smart City



Source: IDC Report – *Building Agile Data Driven Smart Cities,* 2015

A new report by The Wharton School, *Smart Cities: The Economic and Social Value of Building Intelligent Urban Spaces,* suggests that the worldwide trend toward building smart cities is getting the creative juices of urban planners, entrepreneurs, and citizens flowing. Some projects are impressive in their innovation, while others are laudable for their scale and impact. Typically, these projects begin in discrete pockets of existing cities, where they can generate a "proof of concept," improve upon their design while gaining insights, and go on to replicate them in other parts of the city or in other communities. A less common example is the building of brand new cities, such as Songdo in South Korea or one planned in China's Guangdong province. Innovative work from around the world in smart cities touches on a wide range of areas including energy, healthcare, transportation, parking, crime management, adaptive reuse of existing and unused infrastructure, citizen participation, and enhanced overall quality of life.

Other examples of smart cities include Barcelona, where the drive to revolutionize began more than a dozen years ago, Copenhagen with a focus on boosting sustainability, Charlotte, North Carolina, with a successful energy efficiency program, and China's capital city, Beijing. There also is a strong push in the Middle East toward smart cities and adopting IoT and the technologies that benefit from machine learning, originating from the perspective of public safety and security. The U.S., which came later to the game, is now getting its act together at the Federal Government level, although several U.S. cities, including New York, Philadelphia, and San Francisco, launched such initiatives years before the term "smart city" or "digital city" became a buzzword. Smart cities are driven by the amount of data and ability to cross many verticals in the government space.

> Smart cities will represent a market value of USD 1.565 Trillion by 2020, driven by investments in sectors such as smart districts, utilities, healthcare, security, and governance. *– Frost & Sullivan*

In September 2015, the Obama administration announced a smart cities initiative to invest more than $160 million in at least two dozen research and technology collaborations to help communities across the country tackle challenges ranging from fighting crime and reducing traffic congestion to fostering economic growth. Digital cities, based on IoT technologies, use traffic lights to analyze traffic patterns.

According to a report by Frost & Sullivan, *Strategic Opportunity Analysis of the Global Smart City Market*, smart cities will represent a market value of USD 1.565 Trillion by 2020, driven by investments in sectors such as smart districts, utilities, healthcare, security, and governance.

One caveat is that it's harder for the public sector in developed Western economies to cultivate a smart city, with civil liberties concerns centered on if and how data is leaving the city. Security and privacy concerns work to restrict data leaving the city, and more broadly, leaving the host country.

Smart cities also involve putting resources in the right place. As an example, you have limited resources with police, limited staffing for EMTs, limited numbers of fire stations, trucks, ambulances, etc. The goal is to develop predictive models — for instance Austin, TX has intersections that have the most traffic between the hours of 4-6pm, and predicting the most likely intersections to have an accident is a tremendous benefit for public safety. When you look at those areas with traffic issues, patterns emerge and you can allocate resources to those identified parts of the city.
If you can predict that you're going to have five or six accidents in downtown Austin, you can schedule resources around the downtown area in case that happens. As a specific example, a major event like the Austin City Limits (ACL) music festival attracts upwards of 70,000 people in the central part of the city — where do you put emergency services, where do you put police, where do you put traffic cops, where do you put EMTs? Smart cities means allocating resources at the right time.

## Motivations and Challenges for Government Use of IoT

Where federal agencies have begun deploying IoT solutions, it is typical to pursue several primary goals: increase efficiency, reduce costs, and offer new services (e.g. anti-crime surveillance). Major projects include a smart buildings initiative to reduce energy costs, a telematics program to increase the efficiency of government vehicles, an effort to improve asset management, and an automated process to replace manual data-collection.

Several federal agencies have used the IoT as an opportunity to create services in support of their missions. Major new projects that use the IoT include improving national defense, monitoring the natural world, and enhancing safety and public health.

Adoption of IoT technology for government applications, however, doesn't always follow a straight line. The Federal Government faces a number of challenges that have hampered the adoption of the IoT in the public sector. First, there is a lack of strategic leadership at the federal level about how to make use of the IoT. Second, federal agencies do not always have workers with the necessary technical skills to effectively use data generated by the IoT. Third, federal agencies do not have sufficient funding to modernize their IT infrastructure and begin implementing IoT pilot projects. Fourth, even when funding exists, federal procurement policies often make it difficult for agencies to quickly and easily adopt the technology. Finally, risks and uncertainty about privacy, security, interoperability, and return on investment delay federal adoption as potential federal users wait for the technology to mature and others to adopt first.

## Government Sponsored Healthcare and Life Sciences

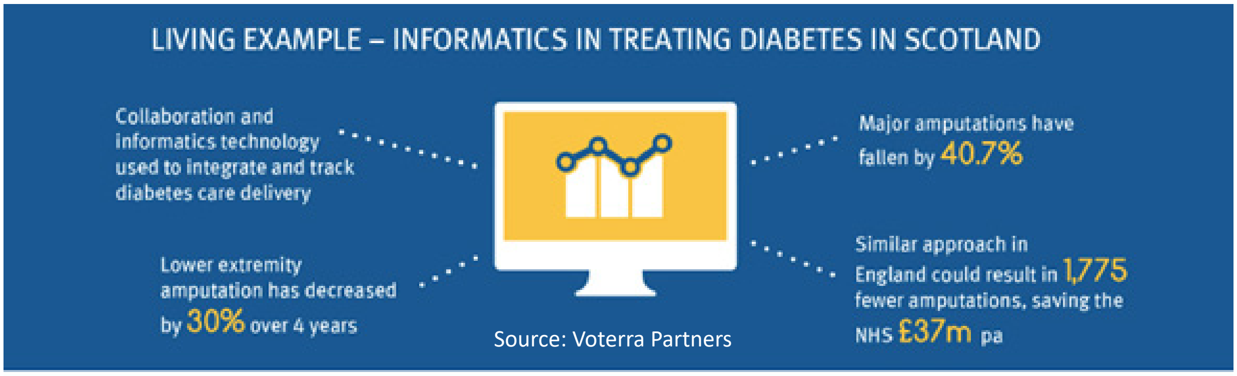
Government sponsored data initiatives within healthcare and life sciences are encouraging — they not only increase transparency but also have the potential to help patients. Not surprisingly, recent years have seen a flurry of activity in this sector in many countries. For example, the Italian Medicines Agency collects and analyzes clinical data on expensive new drugs as part of a national cost-effectiveness program. Based on the results of this effort, the government may reevaluate prices and market access conditions.

Within the U.S., the Federal Government has been encouraging the use of its healthcare data through various policies and initiatives with the hope to directly improve cost, quality, and the overall health-care ecosystem. With more data being released, the Federal Government is trying to ensure that all appropriate stakeholders, including those in private industry, can access the information in standard formats. The HealthData.gov portal, for example, includes federal databases with information on the quality of clinical providers, the latest medical and scientific knowledge, consumer product data, community health performance, government spending data, and many other topics. In addition to publishing information, the aim is to make data easier for developers to use by ensuring that they are machine-readable, downloadable, and accessible via application programming interface (API).

In a report by Voterra Partners for Dell EMC, *Sustaining Universal Healthcare in the UK: Making Better Use of Information*, we learn about the effort to address the unprecedented financial pressure faced by the UK's National Health System — affecting patient's quality of treatment, as waiting times increase, and research funding is restricted. The belief is that the availability of patient information and data analytics could have a substantial beneficial impact. Using big data technology, there is focus on three main areas:

- **Interoperability of patient records –** the ability to access and update records at any point in the healthcare system by integrating NHS institutions.
- **Data analytics –** using large quantities of information to better predict and personalize medicine. Data analytics can identify the combination of factors that put the patient at high risk of developing a chronic condition, allowing the intervention to prevent them from getting ill. Personalized medicine can improve early diagnosis and improve quality of care treatments and outcomes can be analyzed in conjunction with patient details in order to maximize the benefit of any treatment.
- **Mobile technology –** apps can be used by medical practitioners to provide up-to-date practical advice and by individuals to manage their health. Tracking devices are becoming more popular and could be used to maintain personalized healthy lifestyles.

Scotland has used informatics technology to provide an integrated care model for the treatment of diabetes. GPs, patients and secondary care professionals have collaborated to treat diabetes over a period of 20 years. The informatics technology is used to track patients' treatment and treatment outcomes that are carefully monitored and managed so as to reduce the severity of the condition. This has achieved impressive results as shown in the figure provided below.


LIVING EXAMPLE – INFORMATICS IN TREATING DIABETES IN SCOTLAND
Collaboration and informatics technology used to integrate and track diabetes care delivery
Lower extremity amputation has decreased by 30% over 4 years
Major amputations have fallen by 40.7%
Similar approach in England could result in 1,775 fewer amputations, saving the NHS £37m pa

Source: Voterra Partners

## Genomics

Scientists are realizing fascinating new perspectives on the human genome, and it's all thanks to the advancements made in data analytics. For years, genes have been studied and mapped, with perhaps the crowning achievement being the completion of the Human Genome Project in the early 2000s, but true understanding of how human genetics work has required more intensive study and more resources. Only recently have scientists been able to look more closely at human genes, and much of this progress comes as they apply data analytics to the effort.

> The 100K Project is an ambitious program to sequence 100,000 whole genomes from NHS patients across England.

There is much interest in genomics and personalized healthcare across the European Union. In the UK there is the "100K project," where 100,000 genomes are being sequenced. Formally known as "The 100,000 Genomes Project," it is an ambitious program to sequence 100,000 whole genomes from NHS patients across England. Genomics promises significant benefits in healthcare through scientific discovery, and this study will help to deliver on this goal. Dell EMC provides the platform for large-scale analytics in a hybrid cloud model for Genomics England. The project has been using Dell EMC storage for its genomic sequence library, and now it will be leveraging a data lake to securely store data during the sequencing process. Backup services are provided by Dell EMC's Data Domain and Networker.

> Apache Spark's compatibility with the Hadoop platform makes it easy to deploy and support within existing bioinformatics IT infrastructures, and its support for languages such as R, Python, and SQL ease the learning curve for practicing bioinformatics practitioners.

Arizona State University (ASU) worked with Dell EMC to create a powerful high-performance computing (HPC) cluster that supports data analytics. As a result, ASU built a holistic Next Generation Cyber Capability (NGCC) using Dell EMC and Intel technologies that is able to process structured and unstructured data, as well as support diverse biomedical genomics tools and platforms. HPC technology and the Dell EMC Cloudera Apache Hadoop solution, accelerated by Intel, upon which NGCC is based, can handle data sets of more than 300 terabytes of genomic data. In addition, ASU is using the NGCC to understand certain types of cancer by analyzing patients' genetic sequences and mutations.

Apache Spark is an ideal platform for organizing large genomics analysis pipelines and workflows. Its compatibility with the Hadoop platform makes it easy to deploy and support within existing bioinformatics IT infrastructures, and its support for languages such as R, Python, and SQL ease the learning curve for practicing bioinformatics practitioners. Widespread use of Spark for genomics, however, will require adapting and rewriting many of the common methods, tools, and algorithms that are in regular use today.

## Neuroscience

There are many challenges to analyzing neural data. The measurements are indirect, and useful signals must be extracted and transformed in a manner tailored to each experimental technique. Analyses must find patterns of biological interest from the sea of data. An analysis is only as good as the experiment it motivates; the faster we can explore data, the sooner we can generate a hypothesis and move research forward.

One of the reasons neural data analysis is so challenging is that there is no standardization. There are families of workflows, analyses, and algorithms that we use regularly, but it's just a toolbox, and a constantly evolving one. To understand data, we must try many analyses, look at the results, modify at many levels — whether adjusting preprocessing parameters, or developing an entirely new algorithm — and inspect the results again.

With Spark, especially once data is cached, we can get answers to new queries in seconds or minutes, instead of hours or days.

Spark allows the ability to cache a large data set in RAM and repeatedly query it with multiple analyses. This is critical for the exploratory process, and is a key advantage of Spark compared to conventional MapReduce systems. With Spark, especially once data is cached, we can get answers to new queries in seconds or minutes, instead of hours or days. For exploratory data analysis, this is a game changer where the ability to visualize intermediate results is critical.

### BRAIN Initiative

The U.S. based BRAIN Initiative uses big data technologies to map the human brain. By mapping the activity of neurons in the brain, researchers hope to discover fundamental insights into how the mind develops and functions, as well as new ways to address brain trauma and diseases. Researchers plan to build instruments that will monitor the activity of hundreds of thousands and perhaps 1 million neurons, taking 1,000 or more measurements each second. This goal will unleash a torrent of data. A brain observatory that monitors 1 million neurons 1,000 times per second would generate 1 gigabyte of data every second, 4 terabytes each hour, and 100 terabytes per day. Even after compressing the data by a factor of 10, a single advanced brain laboratory would produce 3 petabytes of data annually.

## Infectious Diseases

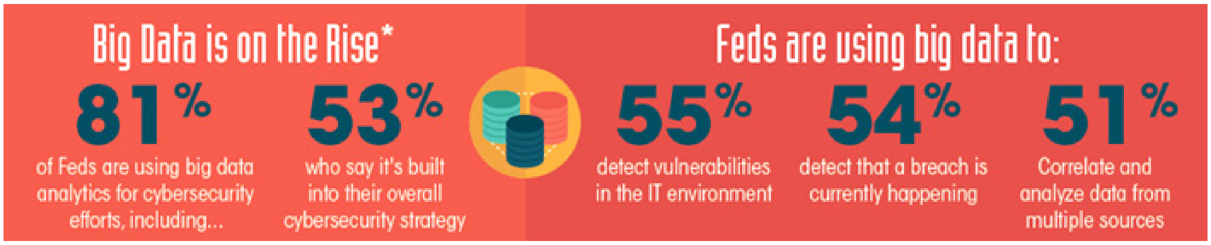
The National Institutes of Health (NIH) is trying to prevent the spread of infectious diseases, e.g. super-viruses. NIH looks at all the drug data and clinical trial results submitted to the FDA, and correlates data from drug manufacturers, doctors, and patients to build a model. As an example, if a super-virus were to take place in a given population, data analytics can help answer questions like: how affected would that population be, how fast would it spread, what actions would the NIH have to take to quarantine that area, and what steps would be needed in order to control the virus before it spreads across the country?

A real-life example is how data analytics is helping the fight against the Zika virus. The World Health Organization has declared the Zika virus a public health emergency that could affect four million people in the next year as it spreads across the Americas. Big data and analytics have played a role in containing previous viral outbreaks such as Ebola, Dengue fever, and seasonal flu, and lessons learned are undoubtedly being put to use in the fight against Zika. However, while statistical modeling of vast, real-time data sets is becoming ingrained across healthcare and emergency response, the support infrastructure needed to put these initiatives to work at ground level is lagging behind.

Data research has dramatically sped up the development of new flu vaccines. By analyzing the results of thousands of tests at institutions around the world, compounds can be developed to target the specific proteins that are found to enable the virus to grow. Big data is also used by epidemiologists to track the spread of outbreaks.

Dell EMC and the University of Cambridge maintain the European HPC Solution Centre. In collaboration with Intel, Dell EMC and Cambridge HPC Solution Centre aims to provide answers to challenges facing the HPC community and feed the results back into the wider research community. Thanks to the Centre, researchers are exploring the genetic analysis of tens of thousands of disease patients.

## Government Use of Big Data for Cybersecurity



Source: Navigating the Cybersecurity Equation by MeriTalk

The cybersecurity waters are teeming with threats by criminals, nation states, and hacktivists, and government agencies do not have the personnel, tools, or time to properly handle the massive amounts of data involved especially with the attack surface constantly expanding. However, with the ability to discover insights in the very data they are sinking in, big data may be the requisite lifeline.

While 90 percent of government data analytics users report they have seen a decline in security breaches, 49 percent of federal agencies say cybersecurity compromises occur at least once a month as a result of an inability to fully analyze data. These are some of the findings in a new report from MeriTalk's (a public-private partnership focused on improving the outcomes of government IT), *Navigating the Cybersecurity Equation*, which examines how agencies are using big data and advanced analytics to better understand cybersecurity trends and mitigate threats.

The survey asked 150 federal cybersecurity professionals to examine how agencies are using big data and analytics to better understand cybersecurity trends and mitigate threats. The study found that 81 percent of federal agencies say they're using data analytics for cybersecurity in some capacity — 53 percent are using it as a part of their overall cybersecurity strategy, and 28 percent are using it in a limited capacity. However, breaches continue to afflict agencies with 59 percent of federal agencies reporting they deal with a cybersecurity compromise at least once a month due to their inability to fully analyze data. In addition, just 45 percent found their efforts to be "highly effective." What is holding agencies back from reaching their cybersecurity goals? Where do agencies go from here?

The top challenges for feds surrounding cybersecurity as reported by participants were:

- The sheer volume of cybersecurity data, which is overwhelming (49 percent)
- The absence of appropriate systems to gather necessary cybersecurity information (33 percent)
- The inability to provide timely information to cybersecurity managers (30 percent)

Because of these challenges, participants stated more than 40 percent of their data goes unanalyzed. Other obstacles include the lack of skilled personnel (40 percent), potential privacy concerns (27 percent), and lack of management support/awareness (26 percent).

And Feds are Struggling to Stay Afloat
Top challenges are:
01
The sheer volume of cybersecurity data is overwhelming (49%)
02
Agencies don't have the right systems in place to gather the cybersecurity information they need (33%)
03
The information is no longer timely when it makes it to cybersecurity managers (30%)

Source: Navigating the Cybersecurity Equation by MeriTalk

> Dell EMC Cyber Solutions Group has been developing applications that enable penetration at a database level with zero impact on the database and provides real-time assessments across the network in seconds. As soon as these threats are identified, assessments can be announced to determine how fit the security solution is.

From a perspective outside the U.S., cybersecurity has been identified as the number one threat to the UK government. The UK is seeing a big emphasis on using *National Cyber Security Center* (NCSC) that was announced early 2016. The UK faces a growing threat of cyberattacks from states, serious crime gangs, hacking groups as well as terrorists. The NCSC will help ensure that the people, public and private sector organizations and the critical national infrastructure of the UK are safer online.

Data analytics is playing a significant role in looking at the threat landscape — to determine where weaknesses are, and whether the right strategies and tools are in place. Additionally, there is a focus between the military and the intelligence services, which are centered on pursuing penetration testing. It is difficult to do penetration testing on live systems, and the challenge is that you'll never really be able to test the vulnerabilities across the network for fear of bringing down critical applications. Dell EMC Cyber Solutions Group has been developing applications that enable penetration at a database level with zero impact on the database and provides real-time assessments across the network in seconds. As soon as these threats are identified, assessments can be announced to determine how fit the security solution is. The Cyber Solutions Group is part of Dell EMC, and features expertise in the realms of advanced storage technologies, high availability, cyber security, big data, cloud computing, and data science.

## Cybersecurity Solutions

Cybersecurity difficulties can be addressed through the use of hardened technology solutions such as the one from SecureWorks, Inc., a subsidiary of Dell Technologies. SecureWorks is a secure, cost-effective, and highly scalable solution for collecting, processing, and analyzing massive amounts of data collected from government agency environments. The solution is deployed on the Dell EMC Cloudera Apache Hadoop Solution, accelerated by Intel (using the Dell EMC Reference Architecture).

### SecureWorks Features

SecureWorks offers a number of features attractive to the needs of government agencies:

- Explosive data growth leading to a search for new technology
- Secure solutions built for big data and analytics
- Innovative security capabilities through an evolving technology stack
- Better protection for government clients
- Faster threat response for government agencies
- Reduced costs and increased scalability

SecureWorks provides proven threat detection. Offered as a service, SecureWorks monitors the environment and looks for any kind of threats whether from the network, or internally, or externally. The reason SecureWorks is based on the Cloudera ApacheTM Hadoop® software platform is because the amount of attacks happening on a given environment is so high. It's necessary to be able to monitor logs from network devices, logs from computers, notebooks, servers, etc. Typically, you don't have anywhere to put all that log data, and you don't have anything fast enough to process and analyze the data. Hadoop is an open source analytics platform, built from the ground up, to address today's big data challenges. It enables government agencies to load and consolidate data from various sources into the highly scalable Hadoop Distributed File System (HDFS). This data can then be processed using highly distributed compute jobs through the MapReduce framework.

By integrating Hadoop with an IT environment, you're able to achieve two important aspects. First, you're able to solve the scalability problem — not in terms of infrastructure, but rather scalability as far as being able to detect a threat — a threat might be a log, a Trojan, something that came in through the firewall. So now you're able to process and analyze that data in any given format and put it in Hadoop. Second, by putting your data in a single repository, like a data lake, you're able to layer on statistical models and algorithms to detect anomalies and detect the threats on top of Hadoop. This enables you to build a single repository to do all your analysis — one way to ingest data, one way to process data, and one way to analyze data — to get an operational model of how you're going to consume all the data and identify a threat.

Another cybersecurity solution is from RSA, also a part of Dell EMC Technologies. RSA security analytics was used by Los Angeles World Airports (LAWA) in order to track everything that happens within its environment. Working frequently with the FBI and the Secret Service, it has to be accountable for its cybersecurity. Its goal was to have real-time detection of security events in order to ensure public safety. RSA security analytics has enabled LAWA to greatly improve the speed of its response to immediate threats. The solution enables deep-dive into payloads before and after a security event and delivers more information about each device than was previously possible. The RSA cybersecurity solution has also helped shorten incident response time as analysts can see all the information in one place, rather than spending time searching for it.

## Case Studies: Dell EMC Focused Customer Use Cases

In order to illustrate how government agencies are rapidly moving forward with the adoption of the big data technology stack, in this section we'll consider a number of use case examples that have benefited from these tools. In addition, these project profiles show how big data is steadily merging with traditional HPC architectures. In each case, significant amounts of data are being collected and analyzed in the pursuit of more streamlined government.

### Dubai's Smart City Initiative

Dell EMC has provided an enterprise hybrid cloud platform to the technology arm of Dubai's smart city initiative. Smart Dubai Government deployed the cloud platform to establish an Information Technology-As-A-Service infrastructure that will work to aid the delivery of new services for individuals and organizations. The cloud platform is designed to provide self-service and automation functions, deeper visibility into performance and capacity usage of infrastructure resources, and a unified backup and recovery system.

The hybrid cloud platform is based on VMware's vCloud Suite and Dell EMC's VPLEX and Vblock converged infrastructure and ViPR Controller software-defined storage automation. The platform was built to integrate scalability, speed, and agility of public cloud platforms with the control and security of private systems.

### San Diego Supercomputer Center

The San Diego Supercomputer Center (SDSC) was established as one of the nation's first supercomputer centers under a cooperative agreement by the National Science Foundation (NSF) in collaboration with UC San Diego. The center opened its doors on November 14, 1985. Today, SDSC provides big data expertise and services to the national research community.

In 2013, SDSC was awarded a $12 million grant by the NSF to deploy the Comet supercomputer with the help of Dell and Intel as a platform for performing big data analytics. Using the Intel® Xeon® Processor E5-2680 v3, SDSC uses parallel software architectures like Hadoop and Spark to perform important research in the field of precision health and medicine. It also performs wildfire analysis and clustering of weather patterns as well as integration of satellite and sensor data to produce fire models in real-time.

### National Center for Supercomputing Applications

Established in 1986 as one of the original sites of the National Science Foundation's Supercomputer Centers Program, the National Center for Supercomputing Applications (NCSA) is supported by the state of Illinois, the University of Illinois, the National Science Foundation, and grants from other federal agencies. The NCSA provides computing, data, networking, and visualization resources and services that help scientists, engineers, and scholars at the University of Illinois at Urbana-Champaign and across the country. The organization manages several supercomputing resources, including the iForge HPC cluster based on Dell EMC and Intel technologies. One particularly compelling scientific research project that's housed in the NCSA building is the Dark Energy Survey (DES), a survey of the Southern sky aimed at understanding the accelerating expansion rate of the universe. The project is based on the iForge cluster and ingests about 1.5 terabytes daily.

### Tulane University

As part of its rebuilding efforts after Hurricane Katrina, Tulane University partnered with Dell EMC and Intel to build a new HPC cluster to enable the analysis of large sets of scientific data. The cluster is essential to power data analytics in support of scientific research in the life sciences and other fields. For example, the school has numerous oncology research projects that involve statistical analysis of large data sets. Tulane also has researchers studying nanotechnology, the manipulation of matter at the molecular level, involving large amounts of data.

Tulane worked with Dell EMC to design a new HPC cluster dubbed Cypress, consisting of 124 Xeon-based PowerEdge C8220X server nodes, connected through the high-density, low-latency Z9500 switch, providing a total computational theoretical peak performance of more than 350 teraflops of computational power. Dell EMC also leveraged their relationship with Intel, who in turn leveraged their relationship with leading Hadoop distribution Cloudera allowing Tulane to do data analytics using Hadoop in an HPC environment.

Using Cypress enables Tulane to conduct new scientific research in fields such as epigenetics (the study of the mechanisms that regulate gene activity), cytometry (the measurements of the existence of certain subsets of cells within a kind of tissue in the human body), primate research, sports-related concussion research, and the mapping of the human brain.

### Translational Genomics Research Institute

To advance health through genomic sequencing and personalized medicine, the Translational Genomics Research Institute (TGen) required a robust, scalable high-performance computing environment complimented with powerful data analytics tools for its Dell EMC Cloudera Apache Hadoop platform, accelerated by Intel. For example, a genome map is required before a personalized treatment for neuroblastoma, a cancer of the nerve cells, can be designed for the patient, but conventional genome mapping takes up to six months. Dell EMC helped TGen reduce the time from six months to less than four hours and enables biopsy to treatment in less than 21 days. Now TGen is widely known for their comprehensive genome analysis and collaboration through efficient and innovative use of their technology to support researchers and clinicians to deliver on the promises of personalized medicine for better patient outcomes.

## Summary

Data analytics for government is a rapidly evolving field, offering exciting opportunities that, when explored and applied, can help fragile states uncover powerful and effective methods for optimizing governance. Furthermore, it is clear that government agencies have an appetite for embracing big data technologies to help transform the public service experience for citizens and employees. Government leaders need to focus on delivering value and on adopting these emerging technologies while creating the kind of internal conditions that will inspire employees to embrace change.