

BEACON REPORT

BIG DATA, BIG BRAINS



For most of human history, we simply did not have enough data. What lay beyond the next hill, across the ocean, and beneath the surface of our skin remained unsolvable because there were no reliable observations to feed the analysis. We were in the dark.

How times have changed. We are no longer starved for data. We are drowning in it.

Mankind created 150 exabytes (billion gigabytes) of data in 2005¹, and 1,800 exabytes in 2011²; growth that only continues to accelerate. Every minute, users:

- Upload 48 hours of video to YouTube
- Send 204 million emails
- Spend \$207,000 via the web
- Create 571 new websites³

Within the Federal government, U.S. drone aircraft sent back 24 years worth of video footage in just 2009. Every 24 hours, NASA's Curiosity rover can send nearly three gigabytes of data, collecting in mere days the equivalent of all human knowledge through the death of Augustus Caesar – from Mars.

And so the puzzle of driving human knowledge forward changes. As the ability to create data expands exponentially, our ability to analyze and convert it to information and knowledge becomes the primary barrier to progress. In fact, the data to cure cancer, unlock the Grand Unified Theory, untangle traffic congestion, and decode consumer behavior may already exist, waiting only to be connected or calculated.

Enter Big Data, the promise of channeling the rising tide of data to needed knowledge. Still much more a concept than a defined set of technical capabilities and user skills, how do we realize that promise? What tools do we need? Where do we get started?

The Big Data, Big Brains

To answer those questions, we queried 17 of the brightest minds from government and industry to define what Big Data is, why it is important, and how to reap the benefits.

A MeriTalk Beacon

This report on Big Data is the first MeriTalk Beacon, a new series of reports designed to shed light and provide direction on far-reaching issues in government and technology. Since Beacons are designed to tackle broad concepts, each Beacon report relies on insight from a small number of big thinkers in the topic area. Less data. More insight. Real knowledge.

The Big Thinkers in Big Data



Tom Soderstrom
NASA



Mike Olson
Cloudera



Bruce Nelson
Oracle



Mike Little
NASA



Paul Gustafson
CSC



Douglas Neal
CSC



Darren Smith
NOAA



Tsengdar Lee
NASA



Karl Horak
Sandia National Labs



Alexander Rossino
Deltek



Jonah Czerwinski
Department of Veterans Affairs



Stanley Tyliczszak
General Dynamics Information Technology



Peter Mell
National Institute of Standards and Technology



Nabajyoti Barkakati
Government Accountability Office



Van Ristau
DLT Solutions



Pete Tseronis
Department of Energy



Mike Giesler
NetApp

What is Big Data?

So what, exactly, is Big Data? The National Institute of Standards and Technology (NIST) – the Federal agency tasked with defining things – defines Big Data as a limit, a line in the sand. Indeed, most of the respondents shared the view of Big Data as the point at which the traditional data management tools and practices no longer apply.

“Big Data is part of an iterative lifecycle that should be part of an over-arching enterprise information strategy.” – *Pete Tseronis, Department of Energy*

More importantly, however, the panel saw beyond Big Data as just a technical limit to add:

Management Task: Big Data is more than a technology issue, it involves processes and training that will be different from the current disciplines of data management and analysis.

Hidden Value: Big Data isn’t just about dealing with larger, faster, and more varied chunks of data. It is about distilling vast data into new, previously unknowable insights.

Peter Mell, Computer Scientist for the National Institute of Standards and Technology defines Big Data as:

“Where the data volume, acquisition velocity, or data representation limits the ability to perform effective analysis using traditional relational approaches or requires the use of significant horizontal scaling for efficient processing.”

Though technology will rightfully dominate the discussion of Big Data, in truth the definition needs to extend beyond just the limits to embrace both the challenge and the ultimate value. Based upon the panel insights, we would amend the NIST definition to:

“Big data is the set of technical capabilities and management processes for converting vast, fast, and varied data into useful knowledge.”



Figure 1: What is Big Data Word Cloud

There's Gold in Them Thar Yottabytes

Like the California Gold Rush, it is clear to prescient data prospectors that the vast exabytes of data contain value – but the question is what value? What should organizational and IT leaders rightly expect from the analysis of vast collections of data? Here, the panel shared four key vectors of insight that Big Data will likely deliver:

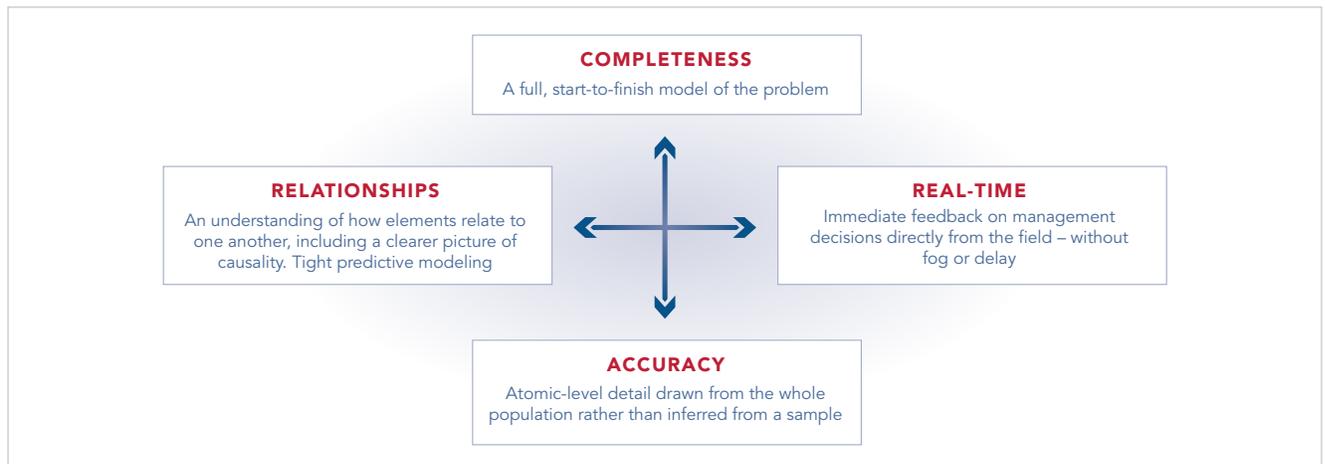


Figure 2: Vectors of Big Data Insight

More importantly for CIOs, the panel also believes that Big Data will provide a Big Opportunity for technologists to play a deeper, more meaningful business role. As real Chief Information Officers, CIOs will own delivery of business-critical insights in consumer behavior, competitive environment, real-time operational effectiveness, and the like. They will prospect for and sell the Big Data gold.

“Data can become a competitive advantage. The opportunity is in analyzing your existing data, combined with new data, in interesting ways to gain competitive advantage, deliver better customer experience, and make more intelligent decisions.” – *Mike Gielser, NetApp*

The Oregon Trail or the Donner Party?

The road to riches is not without risk. Like many investments involving technology, the experts see risk in the potential between buying “Big Data” and benefitting from it. Nearly all of the panelists cited some form of analysis paralysis or risk of getting lost in the review of the data. Other risks included:

Volume vs. Value: Spending too much time and effort looking at the sea of detail and too little time

extracting or understanding the value of the data. “Risks are getting lost in a sea of detail or spending too much time thinking about volume, and not enough time thinking about actual utility or value,” said Mike Olson, Cloudera.

Legal/Privacy: As with most new technology, legal frameworks and precedents are not up to speed with all of the data sources and uses. “It will be a challenge to resolve any legal/policy issues that may be barriers to using the data for analytics,” said Nabajyoti Barkakati, Government Accountability Office.

Data Overconfidence: Data and predictive models are a picture of the real world, and not the real world itself. Overreliance can lead to blind spots in decision making or overlooking a key shift in trends or market drivers. “The downsides include improper planning that silos data unnecessarily and the possibility of drawing incorrect conclusions from analysis,” said Alexander Rossino, Deltek.

The Wrong Data: In all data analysis – big and small – the quality of the output is based on the quality of the inputs. “Garbage in. Garbage out,” said Jonah Czerwinski, Department of Veterans Affairs.

Big Data – Big Hurdles

So the road to Big Data riches is risky. The road is also rocky. The expert panel sees enormous value in successfully harnessing Big Data, but notes that IT organizations must leap a number of hurdles to reap that value:

Personnel: “We need scientists and librarians to curate, index, and mine the data.” – *Tsengdar Lee, NASA*

Qualified IT personnel are already at a premium, and Big Data will require a sizable number of new professionals, both to build IT systems capable of using Big Data and to craft analyses that distill data into insights.

Skill Sets: “More training and more skills for dealing with big data.” – *Mike Olson, Cloudera*

Owing to the scale and open-fronted, open-ended nature of Big Data, it will require new skills to manage. That means new education, new training, and possibly even new certifications specifically for Big Data. Leading universities, such as Northwestern University, are now creating accredited degrees and programs for Data Analytics and Data Scientists.

Data Silos: “Most of the Big Data are behind closed doors. The accessibility is a big problem.” – *Tsengdar Lee, NASA*

Some challenges are new, but others are not. Having the technical ability to tackle Big Data and having the access to data sets across an organization are two very different things. To leverage data in a big way, organizations will need access to it.

Data Ownership: “Access to data belonging to different agencies could be a barrier because there could be legal/policy restrictions on sharing/using data.” – *Nabajyoti Barkakati, Government Accountability Office*

Once data sets are available, data ownership is still a question. Does the use of one data set in a larger analysis confer ownership? What about usage rights and requirements?

Budgets: “Funding and prioritizing the need to solve a real mission-critical problem.” – *Van Ristau, DLT Solutions*

As part of a larger budget, how much investment should Big Data consume? What is a reasonable return, and over what payback period? Why should just IT pay for the capability?



Figure 3: Where to Start

Big Data – Lean Management

As defined by the panel, Big Data involves new management processes for taking full advantage of new data resources. Based upon those responses, the Big Data value chain may look like the following:



Figure 4: The Big Data Value Chain

Importantly, IT will need to play a key role in every step of the Big Data value chain. This will fundamentally change the role of IT within the organization – as big a change for most IT professionals as it will be for the traditional line managers.

Big Data – Big Iron – Big Skills

Without question, Big Data will require changes in how organizations – especially the Federal government – capture and manage data. At the heart of Big Data, however, is an enormous technology challenge – dealing with a data tsunami that will overwhelm organizations not able to take advantage of it. Big Data may hold big promise, but until IT solutions are in place to address the explosive growth of data, it is nothing but a Big Problem.

“Significant technical breakthroughs are needed to meet the rate of data creation and service demands.” – *Darren Smith, NOAA*

At the same time, those technology solutions must directly map to new technologist skill sets. The expert panel addressed tools and skills together, as in the table below.

Requirement	Technology Tools	Technology Skills
Storage and Computing	Cloud Single-volume scaling Distributed systems High Performance Computing Cluster (HPCC)	Data science Platform architecture Data architecture
Network	Bandwidth Low-latency fabric	
Software	Data mining Metadata Visualization	Statistics Data librarian Data visualization artist NoSQL Hadoop/MapR scripts

Figure 5: Big Data Technology Tools and Skills

Big Data Gap Report Findings

- Just 60% of Federal IT professionals say their agency is analyzing the data they collect and less than half (40%) are using their data to make strategic decisions
- Federal IT professionals report, on average, that it will take their agency at least three years to fully take advantage of Big Data
- Agencies estimate they have just 49% of the data storage/access; 46% of the computational power; and 44% of the personnel they need to leverage Big Data and drive mission results

Big Data – Big Start

It is unclear which, if any, agencies are using Big Data today. No respondent indicated that their agency was currently implementing a Big Data initiative, though agencies such as NASA and NOAA indicate that they are reaching the limits of brute force and/or participating in White House working groups on the subject. This is an emerging space, so what should agencies be doing right now to prepare for Big Data?

The People Gap

- Data scientists
- Statisticians
- Training

“Establish a sense of urgency around embracing your data and Big Data functionality. Invest now in data infrastructure that will efficiently store, manage, protect, and scale to meet your Big Data needs.” – *Mike Giesler, NetApp*

1 First, Do No Harm: As data creation continues to accelerate, agencies must invest in the infrastructure to capture and manage that data. Without the necessary computing and storage components in place, no new data will be analyzed and a great deal of it may be lost – forever. “We need to prepare with bigger, faster pipes, vastly larger disk farms, better cyber-security,” said Karl Horak, Sandia National Labs.

2 Tackle Ownership/Sharing: “Policies that permit and encourage sharing of data among agencies will help drive value for government users,” said Mike Olson of Cloudera. Before the stakes get too high, agencies and industry need solutions to difficult data ownership and privacy issues. The time to act is now.

3 Education/Training: Tomorrow’s data scientists and data visualization artists are being educated today – in some other field. The panelists unanimously agree on the need for more specialists in data handling fields, and those programs should start immediately, specifically with the National Science Foundation’s Graduate Research Fellowship program. “The successful data scientists will be technically skilled story tellers. They need to teach the data to tell an interesting story that we didn’t already know,” said Tom Soderstrom, NASA.

4 Identify Partnerships: Panelists underscore that Big Data must be a collaborative effort, citing the GeoData.gov web platform as a key success. The White House Office of Science and Technology Policy is actively assembling agency partnerships through the Big Data Senior Steering Group.

According to NASA’s Tom Soderstrom, “Industry and government should partner both in the approach, tools, and goals.”

5 Try It: Though the full promise of Big Data is yet to emerge, there are already opportunities to leverage unstructured analysis and data visualization into business processes. Panelists recommend that agencies should start trying it today, find opportunities to create value, and then accelerate budgets accordingly. “Pick several data sources and try analysis that you would not have attempted before,” said Douglas Neal, CSC.