Looking for the Right Answers in the Clouds

AFCEA International Cyber Committee: Cloud & Big Data Sub-Committee

(Jamie Dos Santos and Jill Singer)

September 2012

Everyone's talking about Big Data today. Is Big Data a buzz word, real phenomenon, or next evolution in our world? In this AFCEA white paper, we surveyed many notable experts to gain perspectives on Big Data. This paper serves as a primer on Big Data characteristics and provides insights into technology challenges and solutions. The intent is to help federal agencies, companies, and communities develop new solutions for consuming, storing, processing, and analyzing Big Data in order to find the right answers needed to accomplish the mission, gain competitive advantage, and collaborate in more meaningful ways.

It is a BIG DATA world with a current volume of 1.8 zetabytes of data created per year and doubling every two years. Technologies--cloud computing, Hadoop, MapReduce, flash array storage, business intelligence tools, etc-- are vital to an organization's ability to keep pace with Big Data. The Big Data architecture introduces a fourth layer in the cloud computing stack. Knowledge as a Service joins the traditional cloud layers (Infrastructure as a Service, Platform as a Service, and Software as a Service) as a focused layer dedicated to the management and analytics of Big Data including binding concepts such as pedigree, lineage, and provenance of data. Harnessing Big Data will require a combination of technology implementations, business process changes, and workforce training to achieve breakthroughs for your organization.

What is Big Data?

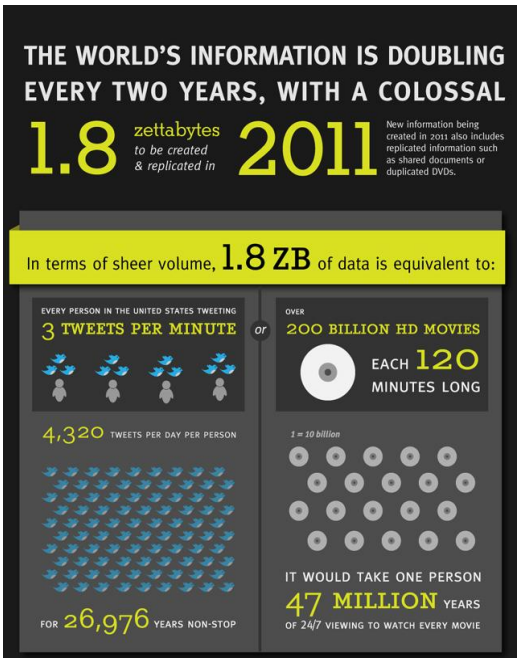
Let's start with a simple definition. McKinsey and Company define Big Data as "Datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze". Big Data can also be characterized by a series of descriptors starting with the letter V. In the 1990s, the word "volume" emerged to describe the rapidly growing rate of data in the Internet age. The table below, from Barbara Wixom of the University of Virginia, offers 15 different data dimensions for Big Data. For brevity's sake, let's examine the four most common descriptors of volume, velocity, variety, and validity.

Data Dimension	Requirements Implications
Volume	Processing, scalability
Variety	Structure, taxonomy
Velocity	Throughput, monitoring
Validity	Quality
Vantage	Stakeholders
Viscosity	Process
Vastness	Visualization
Varied-source	Integration
Value chain	Boundaries
Vicinity	Geospatial
Variant	Behavior
Vorticose	Lineage, networks
Value	Business

Source: Barbara Wixom, 2012

The ***volume*** of data produced in a 24-hour period is staggering and amounts to 1.8 zetabytes per year. Every day, 2 million blogs are posted, 172 million users visit Facebook (spending a combined 4.7 billion minutes on a single social networking site), 51 million minutes of video are uploaded, and 250 million digital photos are shared. We continue to generate 294 billion emails each day, even though many consider email an outdated form of communication.

Perhaps more fascinating is that data ***velocity*** is accelerating. Velocity is the speed at which data is growing and this extreme speed (1.8 zettabytes now and 3.6 zettabytes in 2013) is taxing our current information technology capabilities. According to an IDC Digital Universe Study, we are doubling the world's information every 18 months. This trend will not slow down anytime soon. Will we be able to manage data in our near future? Did you know that each second of high-definition video generates two thousand times as many bytes as one single page of text? IBM research indicates that 90 percent of the world's data has been created in the last two years alone. Apple is selling more iPhones per day than they are babies born in the world, as noted by MBAOnline.com. Samsung's smart phone sales are ahead of Apple sales at an estimated 41 million to 32.6 million for 2Q2012, and demonstrate the rise of additional handheld platforms contributing to the speed at which we create new data.



The third descriptor for Big Data is ***variety*** or the types of data being created. You can generally split variety into structured and unstructured form. The 294 billion emails per day can be considered structured text and one of the simplest forms of Big Data. Financial transactions including movie ticket sales, gasoline sales, restaurant sales, etc., are generally structured and make up a small fraction of the data running around the global networks today. Unstructured data is a primary source of growth in variety. Music is an ever increasing variety of data and we are streaming nearly 19 million hours of music each day over the free music service, Pandora. Spotify, a paid streaming media service, is now the number two revenue source for music labels—second behind Apple's iTunes. Old television shows and movies are another source of variety in the non-structured realm. There are over 864,000 hours of video uploaded to YouTube each day. According to MBAOnline.com, we could pipe 98 years of non-stop cat videos into everyone's home for endless hours of boredom, fun, or insanity!

The biggest challenge from a data variety perspective is harnessing the unstructured information for business relevance and data driven decisions. We've spent decades perfecting analytic tools for structured information. Analytic tools for unstructured data are more limited and less intuitive. Not all relevant marketing data, as an example, is confined to structured business transactions. Tweets, Facebook posts, YouTube video, and so forth, now represent valid indicators to a business. Corporate reputations can be improved or demolished nearly instantaneously by these new sources of data.

Validity is a singular term designed to characterize the quality, pedigree, lineage, provenance, value, integrity, setting, and context for the data. Structured and unstructured data needs validity characterization and it should follow the data from acquisition to retirement. Information from a trusted source is more highly valued than information from a new or casual source. Over time, however, the new source can be further tested and validity of prior data from that source may increase or decrease. Maintaining data heritage is more complex than meta-data tagging and requires situational awareness (i.e., context) when using the data for business decisions.

Is Cloud Computing ready for Big Data? Is IT ready for Big Data?

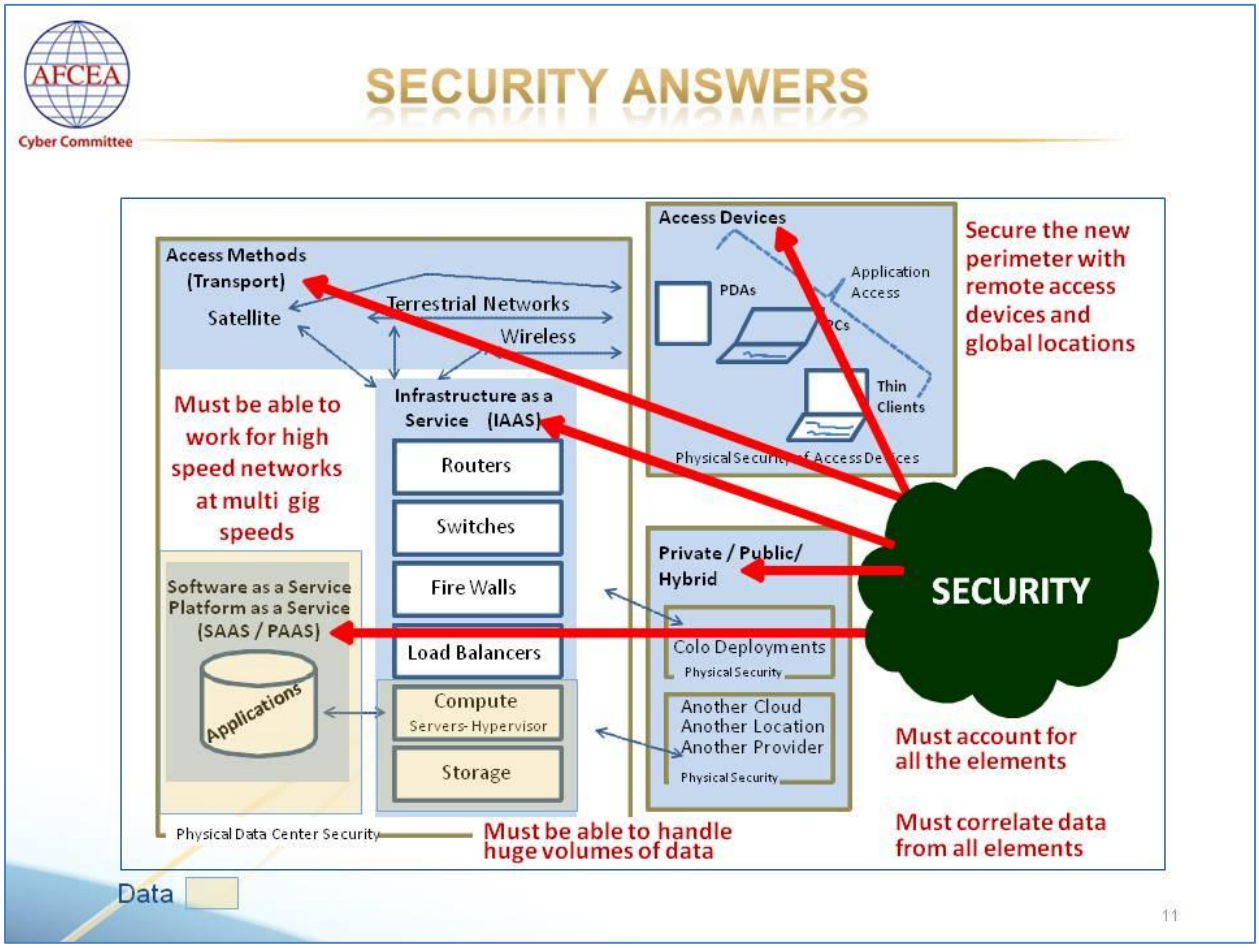
Our survey respondents (notable experts in the Big Data, Cloud Computing, and IT industry) generally concurred on the need for fast, flexible IT infrastructure to support Big Data. Anything that takes the infrastructure challenges out of the way of the business was deemed part of the critical path to success. Tim Estes, Digital Reasoning, noted the cloud "gives you the speed you need". Similarly, Jeff Jonas from IBM declared any infrastructure that lets you "scale up and out affordably is goodness." Four specific areas were identified for deeper investigation as we considered if Cloud Computing and IT departments were ready for Big Data: *security; store and process; sensemaking; and stewardship*.

Security

Security remains the number one obstacle to preventing IT organizations from adopting Cloud Computing. This same AFCEA sub-committee explored security in cloud computing in a white paper released in 2011. The full white paper can be reviewed through this link:

http://www.afcea.org/committees/cyber/documents/cloudcomputingsecuritylessonslearned_FINAL.pdf)

In summary, the 2011 AFCEA white paper noted that cloud computing offers significantly improved visibility and insight that drives new cyber security solutions. Access to cloud computing services in traditional computing environments and in modern mobile environments provides numerous opportunities to gain visibility and retrieve security data points across your infrastructure, platforms, and applications. Collecting pulse points from the high-speed networks used to connect to your cloud provides insight into threats attempting to breach the perimeter of your infrastructure. Remote access devices and global position/location can be detected through other data points, triggering the requirement for additional security access and authorization controls while also providing real-time knowledge of the security status of end-user devices. Constant monitoring of applications and platforms offers additional data collection points for discovering vulnerabilities in applications that can be used to infiltrate the infrastructure. Moreover, merging measures and metrics from co-located environments or other cloud locations in your global enterprise can add yet another layer of data to the collection.

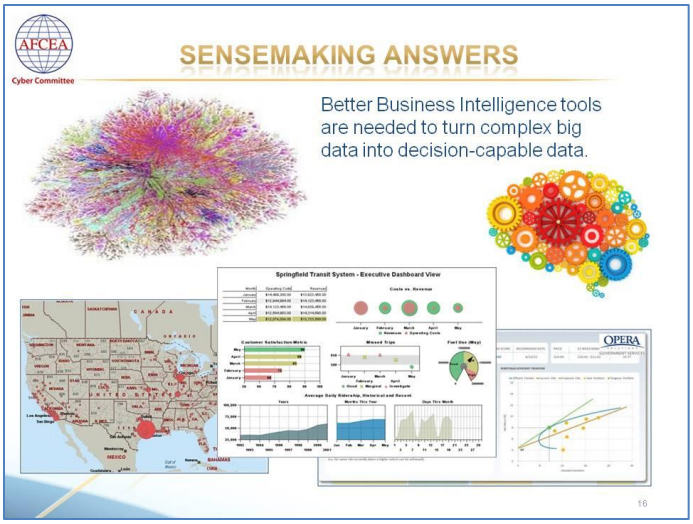


Store and Process

A May 2012 Meritalk Study called "The Big Data Gap" noted that most government IT leaders are fairly positive about the storage and processing resources needed to harness Big Data. The study indicates the respondents currently own about half of the compute resources they actually need. The fear is they

have only about 20 percent of the capacity (in storage and processing power) needed to manage the Big Data headed their way. Industry cloud providers have developed efficient Hadoop cloud-based architectures to handle Big Data. Left on its own, Hadoop will very efficiently process your Big Data but will do so leaving no resources available for other work. Experience indicates restrictions are required on Hadoop jobs to prevent Hadoop from grabbing all available storage and processing resources from the cloud. Industry advances for storage and processing are rapidly emerging to respond to Big Data requirements. Flash array storage is now available at enterprise-class levels from companies like Whiptail and EMC. The InfiniBand trade association created a high performance computing input/output fabric to deliver the internal data center speed needed for Big Data. Gaming providers like NVidia have implemented high-speed clouds for acceleration using graphical processing units (GPU). Moreover, global network providers continue to drive performance advances in optical and electrical equipment with a goal of keeping pace with the explosion of streaming bits and bytes. Cisco's Global Forecast predicts global network traffic will exceed 110 Exabytes per month by 2016. More answers and technologies are emerging each day to overcome remaining deficiencies in storage and processing capabilities for Big Data.

Sensemaking



In the world of Big Data, making sense of the data is not trivial. The volume, velocity, variety, and validity of the data now available can create link analysis diagrams more closely resembling nature's most intricate floral design. You would need Sheldon Cooper's Big Bang Theory eidetic memory to begin to make sense of the data…not to mention a movie theater-sized computer monitor screen! What if your primary viewing device is a smart phone or a 5 inch tablet? Both Gartner and Forrester predict significant increases in the use of mobile devices as Business Intelligence access platforms.

What you need are solutions that make sense of new data in time to derive new observations as the observations actually happen. You also need solutions that allow you to make decisions fast enough to do something about the old and new observations while the transaction is still happening. The tools must protect various slices of the complex data to ensure the user is approved to see the underlying data sources based on the provenance of each individual data element. Advanced analytic and security tools to separate data appropriately, whether your data is in a public cloud or a private cloud, are necessary. Commercial cloud providers are keenly focused on data and privacy protection to ensure only authorized users gain access--using technologies such as encryption, identity management, authorization services, etc. Commercial cloud providers are accumulating very successful track records for safeguarding information.

Stewardship

The fourth challenge for the IT department is data stewardship—often considered data ownership and/or data management. The Big Data Gap MeriTalk survey from May 2012 indicated that nearly 70 percent of respondents thought the "IT Department" had some role in owning and managing the data. Less than 30 percent of respondents indicated the department generating the data should have these roles. Regardless of the decision, someone has to own and manage the data.

The link between financial performance and effective data management is strengthening as companies learn to harvest Big Data. The Economist Intelligence Unit indicates strategies for collecting and analyzing data need to rise to the C-suite level. In essence, Big Data makes IT even more strategic to the business. New/modified career tracks are emerging (in forward-leaning organizations) to develop employees with the skills necessary to validate data sources, define and measure pedigree and lineage of data sources, and effectively manage Big Data for an enterprise.

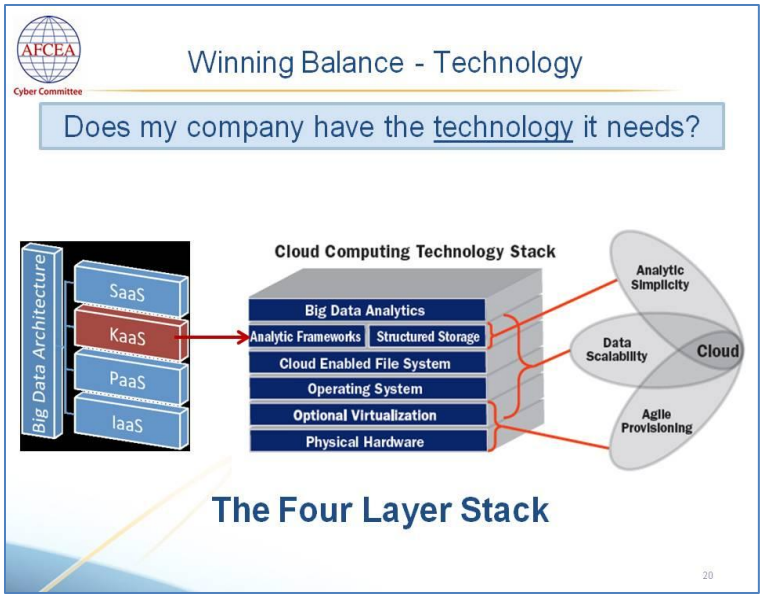
Winning in a Big Data World

As noted above, Big Data is complex and the challenges it presents are daunting. In today's environment, Big Data demands phenomenal corporate balance. Success and competitive advantage require you to focus on technologies, business processes, and people. Your company needs:

- New technologies for controlling Big Data;
- Business processes designed for rapid decisions using Big Data; and
- People trained to make smart decisions exploiting Big Data.

Technology

The Cloud Computing technology stack is evolving to handle Big Data. Our survey revealed the need for a new, four layer stack associated with a Big Data architecture. The Cloud stack of Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS) must shift to accommodate the introduction of Knowledge as a Service (KaaS) in between PaaS and SaaS. This figure depicts the insertion of the new KaaS layer to drive Big Data results. Without a Knowledge layer, companies end up investing heavily in customer knowledge engineering, adaptors, and connectors--essentially ending the elastic advantage of the Cloud at the IaaS layer. This slows

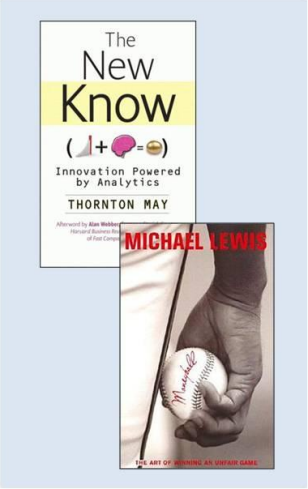


application development and requires human-intensive work on data maps/models and ontologies across disparate systems. The KaaS layer pushes you into a heavily automated, algorithm-driven common knowledge layer that embraces Cloud speed and elasticity.

Innovators have embraced this KaaS idea with Google's PageRank algorithm as an example of a common index with searchable attributes capable of working across the diversity of the Web without relying on substantial amounts of manual organization. Deep insights from Big Data require semantic silos—both structured and unstructured content--to be "knowledge processed" and moved into the Cloud in order to capitalize on the business advantages opportunities of data mining. The specific technologies to ingest and hold structured and unstructured data (as repositories) are similar with any processing differences derived from the complexity of business needs. The repositories are living; as new data arrives, its connection to existing data will be identified and may new data may change the value of the old data to your business. This concept of "data finding data" is the subject of IBM efforts under the direction of Jeff Jonas. You can read more at the below site.

http://jeffjonas.typepad.com/jeff_jonas/2009/07/data-finds-data.html

Business Processes



In a Big Data world, companies should pay as much attention to the data as they do their other corporate assets, e.g., labor and capital. Company goals must be clear with labor, capital, and data aligned. Your employees should be empowered to make decisions with proper checks, balances, and audits built in. Speed is critical and using multi-layered and lengthy paths to finalize decisions will limit your competitive advantage. Business processes designed around lessons learned and adaptability will facilitate Big Data organizations. Governance models and decision thresholds for employees should be clear, with escalation paths obvious and some understanding of the potential "mosaic effect" present. The "mosaic effect" occurs when seemingly unclassified or benign data are combined together by an analyst with a resulting picture that becomes confidential or more highly classified. Training your employees to recognize the "mosaic effect" is necessary to protect sensitive results, intellectual property, and competitive advantage. Finally, adhering to strict data management rules (e.g., process once and use many times), will facilitate long-term Big Data utility by protecting the integrity and provenance of your corporate data.

People

The third element in your winning balance in a Big Data world is your people. Bad decisions in today's world are more obvious and less tolerated. Sometimes, analysis of Big Data drives a company to the wrong conclusion or decision. Other times, Big Data can be successfully used to back up instincts with

facts. Big Data in financial institutions has become so complex, the Federal Deposit Insurance Corporation created an entire office for overseeing Complex Financial Institutions (FDIC/CFI).

Companies that successfully train their people to use Big Data wisely will reap financial and market share rewards. Alternative analysis, critical thinking, and other analytic skills—combined with experiential learning and mentoring--will be necessary to ensure your team is seeing the right answer in the possibilities. Thornton May's book, The New Know, highlights the requirement for companies to capitalize on data and brain power to have good knowledge of what happens next…not what happened in the past. Generally speaking, everybody will have a Moneyball team in the future. What's going to set your team apart?

Big Data Summary

Cloud Computing provides the technology foundation to capitalize on Big Data for corporate success. The flexible infrastructure offered through Cloud Computing--combined with increased storage and processing power of new technologies in the Cloud—provide the rich, agile compute platform to handle the volume, variety, velocity, and validity needs of Big Data. Insert a fourth layer into the Cloud Computing stack (Knowledge as a Service) between PaaS and Saas to reduce human-intensive ontology work in favor of automated, algorithm-driven features designed to exploit disparate data in the Cloud.

Beyond technology, Big Data is likely to require changes in your business processes to ensure decisions with proper analytic judgment with necessary oversight operating at the right speed for competitive advantage. Spend time training your people to analyze data from alternative points of view and to quickly accept automatically-generated observations. People make decisions; data doesn't. Drowning a poorly trained employee in loads of data will still produce poor decisions.

Social media, instantaneous access, and an "always connected" population of stakeholders will increasingly demand transparent accountability. Use Big Data through Cloud Computing to demonstrate your corporate decisions are clearly backed by facts.

Sources

Cloud and Big Data Experts Surveyed

- Zalmai Azmi, Senior Vice President, CACI International, Inc.
- Charles Croom, Vice President, Lockheed Martin Information Technology
- Christopher Day, Senior Vice President, Terremark Federal Group (A Verizon Company)
- John Dvorak, Federal Bureau of Investigation
- Tim Estes, Chairman and CEO, Digital Reasoning Systems
- Jeff Jonas, Chief Scientist, IBM
- Barbara Wixon, Associate Professor of Commerce, University of Virginia

Other Sources Cited and/or Used

- How Much Data Will Humans Create & Store This Year, Josh Catone (June 28, 2011)
- Three Big WHATs to Identify Big Data Challenges, Pearl Zhu (April 2012)
- Bringing big data to the enterprise, IBM
- The BIG Picture on BIG DATA, GovConExec (April 2012)
- Big Data: The next frontier for innovation, competition, and productivity, McKinsey Global Institute (May 2011)
- Big Data: Harnessing a game-changing asset, The Economist Intelligence Unit (September 2011)
- A Day in the Life of the Internet, Matt Silverman (March 06, 2012)
- How to Be Ready for Big Data, Thor Olavsrud (March 20, 2012)
- The Big Data Gap, Meritalk (May 7, 2012)
- Cloudy with a Chance of Savings, Meritalk (April 25, 2012)
- The New Know, Thornton May
- Moneyball, Michael Lewis