

By The Numbers

Big data has largely been driven by the exploding growth of the Internet. Industry watchers estimate the number of connections to the Internet surpassed the global population in 2008. Projections for 2020 are expected to exceed 50 billion connections.

The McKinsey Global Institute has conducted extensive research on big data, and estimated that organizations globally stored more than seven exabytes of new data on disk drives in 2010, while consumers stored more than 6 exabytes of new data on PCs and notebook computers. One exabyte of data is more than four times the amount of data stored in the U.S. Library of Congress, McKinsey has reported. In federal government organizations, market research firm IDC reports data is doubling in size every other year with no slowing in sight.

Many organizations in both the public and private sectors face challenges processing the amount of big data produced from sensors and devices, for example, let alone gaining operational value from that data. What makes 'big data' truly big is a question that's not easily answered. According to Ashit Talukder, Information Access Division Chief in NIST's Information Technology Laboratory, big data is difficult to capture, store, search, share and analyze. And it's growing fast. "Big data may include billions to trillions of records, that are loosely structured or often unstructured," Talukder said.

The records are largely heterogeneous and multimodal, and may be distributed across a number of networks and/or cloud environments, Talukder continued. The records also tend to have complex interrelationships and come from a wide variety of diverse sources.

INTRODUCTION

Depending on how it's viewed, big data presents either an enormous headache, or an amazing opportunity. The ability to sift through massive amounts of structured and unstructured data to reveal useful facts in real-time may help government organizations make better decisions that streamline operations and refine constituent services.

Meanwhile, a recent survey from AIIM reveals that more than 60% of IT executives would find it "very useful" to be able to link structured and unstructured datasets. And more than half of the respondents in the same survey said they would find it "very valuable" (56%) or "hugely valuable" (18%) to be able to carry out sophisticated analytics on unstructured data.

The full report, "Big Data – Extracting value from your digital landfills" can be downloaded from the AIIM website at www.aiim.org/Research/Industry-Watch/Big-Data-2012.

PRIMARY ATTRIBUTES OF BIG DATA

Big data is typically characterized by the following key properties:

- Volume – a massive data size;
- Velocity – a fast rate of data flow into the organization;
- Variety – a range of heterogeneous types of data, networks, nodes are involved.

In addition, when used to refer to an approach, big data typically refers to methods for data-enabled discovery in which the amount of available data or the ability to use data in unique combinations enables discoveries not possible with other approaches.

Currently, only a small portion of the data collected in government organizations is processed and analyzed. According to Ashit Talukder, Information Access Division Chief in NIST's Information Technology Laboratory, the volume and complexity of big data poses many challenges. However, big data also has great potential for 'knowledge-driven' analysis and discovery, rather than 'hypothesis-driven' discovery. "It enables the potential for solving previously unsolved problems, and making new discoveries from previously unprocessed data in a number of different domains," he said.

A Big Data Definition

While no universal definition exists, big data refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze. This definition can vary by sector, depending on what kinds of software tools are commonly available and what sizes of datasets are common in a particular domain. Source: National Institute of Standards and Technology (NIST)

Big Data in the Clouds

Cloud computing delivers an optimized storage, computing, access and visualization environment for big data. According to NIST, cloud interoperability will potentially enable large datasets residing in different clouds to be interoperable with each other, increasing the ability to share, collaborate and analyze multiple very large datasets for knowledge-driven research.

Cloud computing creates a unique opportunity to host, store, process and access big data in a scalable manner that enables elastic, on-demand, anytime-anywhere access from any platform, said Ashit Talukder, Information Access Division Chief in NIST's Information Technology Laboratory.

According to Talukder, government institutions may soon be able to leverage cloud computing for big data challenges, to:

- Provide millions of researchers with unprecedented access to powerful tools;
- Enable a massive shortening of cycle times in time-consuming research processes;
- Reduce research IT costs dramatically via economies of scale.

Despite the potential advantages of cloud computing for helping organizations to analyze the flow of big data, Talukder maintains there are still many elements that must be improved to turn the promise of big data analytics into reality. For example, there's a need for better standards, metrics and interoperability of big data software, algorithms, hardware and infrastructure, he explained. "Advances in fundamental mathematics and statistics are needed, including machine learning for big data, analytics and pattern recognition for big data, subsampling and uncertainty metrics," he said.

Talukder also cited a strong requirement for algorithmic advances in handling massive and complex data, along with better visualization and usability, better clustering, classification, outlier detection, security and privacy for big data. Meanwhile, technological improvements in networking, hardware and software infrastructure for big data storage, computation, and display/visualization are also needed, he explained.

EMERGING USES OF BIG DATA FOR ANALYTICS

Big data analytics is likely to be deployed in many fields to address key operational process challenges and unearth new discoveries based on empirical evidence that results from using actual data, rather than other more traditional forms of analysis. Some of the leading industries and areas that may gain benefits from big data analytics include:

- Environmental and earth sciences
- Medical science
- Astronomy
- Cyber security
- Forensics (both physical and computer/network forensics)
- Fraud detection
- Social media analytics
- Complex network systems design and operation
- Logistics optimization for transportation
- Intellectual property management
- Weather forecasting
- Natural resource exploration and conservation
- Predictive damage assessments in the aftermath of disasters

Research from Stamford, Conn.-based Gartner Inc., underscores a shift to greater 'context-aware' security, for example, in which data from sensors and all devices on the network can be used to aid in defending against threats utilizing modeling and analytics, even when other current security tools for authentication have previously deemed a transaction as safe. Gartner predicts that big data analytics will likely make it possible to increase monitoring, to help organizations of all kinds to better compensate for their loss of direct control over data and systems when implementing cloud-based services.

Cloud Computing and Big Data Work Well Together

CLOUD PROVIDES	BIG DATA NEEDS
On demand self-service	Fault tolerance
Ubiquitous network access	Multiple-protocols
Resource pooling	Scalability (storage, memory, network, etc.)
Rapid elasticity	Scalability (nodes allocation/teardown)
Hybrid (public and private) Cloud with Restricted Access	Secure data access

Source: NIST

Big Data Requires Big Thinking and Intelligent Solutions

It is estimated that U.S. government agencies will add a full Exabyte of data to their data stores during the next two years—the equivalent of over 62 million 16 GB iPads!

When the rate of data growth is combined with the velocity or bandwidth required to move all that data, much of it unstructured data created by video, audio, social media and other means, the issue becomes clear. The problem is one of Big Data—data sets whose size and complexity is beyond the ability of standard tools to capture, store, manage and analyze within a tolerable elapsed time.

“Organizations are at an inflexion point with respect to their data. It will become difficult to do business as usual,” says Dale Wickizer, Chief Technology Officer for the U.S. Public Sector at NetApp, a leader in storage and data management. “If something doesn’t change, the data will bury you and become a huge cost and risk burden to the infrastructure and the mission. But if you figure out how to harness it, it can become an asset.”

A recent MeriTalk study bears this out. Overwhelmingly, agency executives want better ways to unlock and harness agency data to improve efficiency, speed decision-making, and improve forecasting ability. Agencies estimate that today, they have just 49% of the data storage and access, 46% of the computational power and 44% of the personnel they need to leverage Big Data and drive mission results.

The game changer

NetApp has a suite of solutions designed to address all aspects of Big Data, including high performance computing (HPC), based upon NetApp’s modular E-Series storage product. E-Series was designed with fast performance, physical density

and scalability in mind. It can handle large, complex datasets that involve multiple storage systems without compromising data protection and integrity.

Government agencies are taking note. The Energy Department’s Sequoia Project at Lawrence Livermore National Lab is the fastest Blue Gene supercomputer in the world. It is capable of achieving 20 PetaFLOPS of peak performance, using 1.6 million processor cores, and 55 Petabytes of NetApp E-Series storage. That storage system will provide more than 1 Terabyte per second of write performance to the disk subsystem.

Large HPC environments like Sequoia can literally run tens of thousands of high capacity disk drives. When high capacity disks fail, rebuilds can take a very long time and degrade system performance. To address this concern, the latest version of the E-Series Operating System (EOS) supports the use of Dynamic Disk Pools, an innovation that essentially virtualizes RAID protection and dramatically speeds the recovery of high capacity drives (from days, down to a few hours).

Another NetApp technology highly useful for Big Data environments is the more familiar FAS Series of enterprise storage systems. It combines intelligent caching, integrated data protection, storage efficiency features, non-disruptive operations, and massive scale to create a multi-protocol, agile data infrastructure that can support the most demanding cloud and virtualized environments.

For organizations with large, geographically distributed data repositories, NetApp’s StorageGRID® software and E-Series storage is often the best route. StorageGRID is an object-based technology, which can support billions of objects in a

With Big Data becoming a bigger storage challenge each year, it makes sense to engage with professionals who understand storage and federal government technology challenges. CDW’s knowledge of government procurement and dedicated agency account managers, combined with NetApp technology, is a recipe for success.

For more information visit CDWG.com/netapp

single object space, across numerous geographically dispersed sites. The latest version of StorageGRID supports the SNIA Cloud Data Management Interface (CDMI) for accessing data using open standards. Objects are distributed in accordance with a powerful policy engine, allowing those objects to be placed on different tiers of disk or tape, and at one or more locations to balance cost and protection. Strong error checking and correction at the software layer and in the storage guard against “bit rot.” For even stronger protection, objects can be encrypted. •

To learn more about the solutions NetApp offers, please visit netapp.com/bigdata.



To learn more about the CDW-G and NetApp partnership, go to cdwg.com/netapp.

