The Democratization of Big Data

Sean Fahey*

In recent years, it has become common for discussions about managing and analyzing information to reference "data scientists" using "the cloud" to analyze "big data." Indeed these terms have become so ubiquitous in discussions of data processing that they are covered in popular comic strips like Dilbert and the terms are tracked on Gartner's Hype cycle.[1] The Harvard Business Review even labeled data scientist as "the sexiest job of the 21st century."[2] The goal of this paper is to demystify these terms and, in doing so, provide a sound technical basis for exploring the policy challenges of analyzing large stores of information for national security purposes.

It is worth beginning by proposing a working definition for these terms before exploring them in more detail. One can spend much time and effort developing firm definitions for these terms – it took the National Institutes of Science and technology several years and sixteen versions to build consensus around the definition of cloud computing in NIST Special Publication 800-145[3] – the purpose here is to provide definitions that will be useful in furthering discussions of policy implications.

Rather than defining big data in absolute terms (a task made nearly impossible by the rapid pace of advancements in computing technologies) one can define big data as a collection of data that is so large that it exceeds one's capacity to process it in an acceptable amount of time with available tools. This difficulty in processing can be a result of the data's volume (e.g., its size as measured in petabytes[4]), its velocity (e.g., the number of new data elements added each second), or its variety (e.g., the mix of different types of data including structured and unstructured text, images, videos, etc . . .).[5]

Examples abound in the commercial and scientific arenas of systems managing massive quantities of data. YouTube users upload over one hundred hours of video every minute,[6] Wal-Mart processes more than one million transactions each hour, and Facebook stores, accesses and analyzes more than thirty petabytes

---

* DHS Programs Manager, Applied Physics Lab, and Vice Provost for Institutional Research, The Johns Hopkins University.

1. Scott Adams, *Dilbert*, Dilbert (July 29, 2012), http://dilbert.com/strips/comic/2012-07-29/; *Gartner's 2013 Hype Cycle for Emerging Technologies Maps Out Evolving Relationship Between Humans and Machines*, Gartner (Aug. 19, 2013), http://www.gartner.com/newsroom/id/2575515.

2. Thomas H. Davenport & D. J. Patil, *Data Scientist: The Sexiest Job of the 21st Century*, Harv. Bus. Rev., Oct. 2012, at 70.

3. Nat'l Inst. of Standards & Tech, *Final Version of NIST Cloud Computing Definition Published*, NIST (Oct. 25, 2011), http://www.nist.gov/itl/csd/cloud-102511.cfm.

4. One petabyte is equal to one million gigabytes.

5. Edd Dumbill, *What is Big Data*, O'Reilly (Jan. 11, 2012), http://strata.oreilly.com/2012/01/what-is-big-data.html.

6. *Statistics*, YouTube, http://www.youtube.com/yt/press/statistics.html.

of user-generated data.[7] In scientific applications, the Large Hadron Collider generates more than fifteen petabytes of data annually which are analyzed in the search for new subatomic particles.[8] Looking out into space rather than inward into particles, the Sloan Digital Sky Survey mapped more than a quarter of the sky gathering measurements for more than 500 million stars and galaxies.[9]

In the national security environment, increasingly high quality video and photo sensors on unmanned aerial vehicles (UAVs) are generating massive quantities of imagery for analysts to sift through to find and analyze targets. For homeland security, the recent Boston marathon bombing investigation proved both the challenge and potential utility of being able to quickly sift through large volumes of video data to find a suspect of interest.

While the scale of the data being collected and analyzed might be new, the challenge of finding ways to analyze large datasets is a problem that has been around for at least a century. The modern era of data processing could be considered to start with the 1890 census where the advent of punch card technology allowed the decennial census to be completed in one rather than eight years.[10] World War II spurred the development of code breaking and target tracking computers, which further advanced the state of practice in rapidly analyzing and acting upon large volumes of data.[11] The Cold War along with commercial interests, further fueled demand for increasingly high performance computers that could solve problems ranging from fluid dynamics and weather to space science, stock trading and cryptography.

For decades the United States government has funded research to accelerate the development of high performance computing systems that could address these challenges. During the 1970s and 1980s this investment yielded the development and maturation of supercomputers built around specialized hardware and software (e.g., Cray 1). In the 1990s, a new generation of high performance computers emerged based not on specialized hardware but instead clustering of mass-market commodity PCs (e.g., Beowulf clusters).[12] This cluster computing approach sought to achieve high performance without the need for, and cost of, specialized hardware.

The development and growth of the internet in the 1990s and 2000s led to the development of a new wave of companies including Google, Amazon, Yahoo and Facebook that captured and needed to analyze data on a scale that had been previously infeasible. These companies sought to understand the relationships among pieces of data such as the links between webpages or the purchasing

7. *A Comprehensive List of Big Data Statistics*, WIKIBON BLOG (Aug. 1, 2012), http://wikibon.org/blog/big-data-statistics/.

8. *Computing*, CERN, http://home.web.cern.ch/about/computing.

9. *Sloan Digital Sky Survey*, SDSS, http://www.sdss.org/.

10. Uri Friedman, *Anthropology of an Idea: Big Data*, FOREIGN POLICY, Nov. 2012, at 30, 30.

11. *Id.*

12. *See* Thomas Sterling & Daniel Savarese, *A Coming of Age for Beowulf-Class Computing*, *in* 1685 LECTURE NOTES IN COMPUTER SCIENCE: EURO-PAR '99 PARALLEL PROCESSING PROCEEDINGS 78 (1999).

patterns of individuals, and use that knowledge to drive their businesses. Google's quest for a better search engine led their index of web pages to grow from one billion pages in June 2000 to three billion in December 2001 to eight billion in November 2004.[13] This demand for massive scale data processing led to a revolution in data and the birth of the modern "big data" era. These companies developed scalable approaches to managing data that built upon five trends in computing and business. The result of their efforts, in addition to the development of several highly profitable internet companies, was the democratization of big data analysis – a shift which resulted in big data analysis being available not only to those who had the money to afford a supercomputer, or the technical skill to develop, program and maintain a Beowulf cluster, but instead to a much wider audience.

The democratization of big data analysis was possible because of multiple independent ongoing advances in the computing including the evolution of computer hardware, new computer architectures, new operating software and programming languages, the open source community, and new business models.

The growth in computing capability begins with advances at the chip and storage device level. For nearly fifty years, the semiconductor industry has found ways to consistently deliver on Gordon Moore's 1965 prediction that the number of components on an integrated circuit would double every year or two. This has led to the development of increasingly powerful computer chips and, by extension, computers. At the same time, manufacturers of storage devices like hard drives have been able to also achieve exponential increases in the density of storage along with exponential decreases in the cost of storage.[14] "Since the introduction of the disk drive in 1956, the density of information it can record has swelled from a paltry 2,000 bits to 100 billion bits (gigabits), all crowded in the small space of a square inch. That represents a 50-million-fold increase."[15] This increase in computing power and storage capacity has outstripped the needs of most individuals and programs and led to a second key innovation underlying modern big data – virtualization.

Virtualization is the process of using one physical computer to host multiple virtual computers. Each virtual computer operates as though it were its own physical computer even though it is actually running on shared or simulated hardware. For example, rather than having five web servers each operating at 10% capacity, one could run all five web servers on one virtualized server operating at 50% capacity. This shift from physical to virtual machines, has created the ability to easily add new storage or processing capabilities on demand when needed and to modify the arrangement of those computers virtually rather than having to physically run new network cabling. This has led

13. *Our History in Depth*, GOOGLE, http://www.google.com/about/company/history/.

14. Chip Walters, *Kryder's Law*, SCI. AM., Aug. 2005, at 32; *see also* Matt Komorowski, *A History of Storage Cost*, MKOMO BLOG, http://www.mkomo.com/cost-per-gigabyte (graphing the decrease in hard drive cost per gigabyte over time).

15. Walters, *supra* note 16, at 32.

to the growth of "cloud computing" which NIST defines as "a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction."[16] Providers such as Amazon have created services (e.g., Amazon Web Services) to allow individuals and companies to benefit from cloud computing by purchasing storage and processing as required. Rather than investing up front in a datacenter and computing hardware, companies can now purchase computing resources as a utility when needed for their business.

While increased computing power, virtualization, and the development of cloud computing business models were fundamental to the advent of the current big data era, they were not sufficient. As late as 2002, analysis of large quantities of data still required specialized supercomputing or expensive enterprise database hardware. Advances in cluster computing showed promise but had not yet been brought to full commercial potential. Google changed this between 2003 and 2006 with the publication of three seminal papers that together laid the foundation for the current era of big data.

Google was conceived from its founding to be a massively scalable as a web search company.[17] It needed to be able to index billions of webpages and analyze the connections between those pages. That required finding new web pages, copying their contents onto Google servers, identifying the content of the pages, and divining the degree of authority of a page. The PageRank algorithm developed by Larry Page laid out a mathematical approach to indexing the web but required a robust information backbone to allow scaling to internet size.

Google's first step to addressing this scalability challenge was, around 2000, to commit to using commodity computer hardware rather than specialized computer hardware. In doing so the company assumed failures of computers and disk drives would be the norm and so had to design a file system to have constant monitoring, error detection, fault tolerance, and automatic recovery. Google developed a distributed file system (called the Google File System) that accomplished this by replicating data across multiple computers so that the data would not be lost if any one computer or hard drive were to fail. The Google File System managed the process of copying and maintaining awareness of file copies in the system so programmers didn't have to.[18]

---

16. NAT'L INST. OF STANDARDS & TECH, THE NIST DEFINITION OF CLOUD COMPUTING, SPECIAL PUBLICATION 800-145, at 2 (2011), *available at* http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf.

17. Sergey Brin & Lawrence Page, *The Anatomy of a Large-Scale Hypertextual Web Search Engine*, COMP. NETWORKS & ISDN SYS., Apr. 1998, at 107, *available at* http://infolab.stanford.edu/backrub/google.html.

18. SANJAY GHEMAWAT, HOWARD GOBIOFF, & SHUN-TAK LEUNG, THE GOOGLE FILE SYSTEM (2003), *available at* http://static.googleusercontent.com/media/research.google.com/en/us/archive/gfs-sosp2003.pdf.

Google second innovation to address scalability was to develop a new programming language to allow processing of the data in the distributed file system. Traditionally, scalability of high performance computing systems was limited by the ability to move large quantities of data from its storage location to the computer's processor. Recognizing this limitation, Google adopted a parallel computing paradigm and pushed the (much smaller) program out to where the data was stored rather than moving the data to a processing node. Google developed a programming model called Map/Reduce that allowed for the development of programs that could run across the distributed data in the Google File System.[19]

Finally, Google developed a large-scale distributed database called Big Table that could store and analyze data at scales that exceeded the capability of conventional relational databases.[20] This final piece, in conjunction with the Google File System and Map/Reduce provided Google with the infrastructure not only to provide search of indexed webpages but also to provide other offerings at massive scale such as Google Earth and Google Analytics.

While these innovations were profound and highly profitable for Google, their publication as a series of papers from 2003 to 2006 led directly to the democratization of big data analysis. The papers were instrumental in guiding the thinking and development of other technology innovators and companies such as Yahoo! and Facebook. One innovator, Doug Cutting, adapted the concepts in the Google papers into an open source search engine. Later with the support of Yahoo!, this project became the open source Hadoop project with the Apache Software Foundation.[21] Since its launch, Hadoop has become a common tool for analysis of big data and is used in many web companies (e.g., Facebook, eBay, Twitter, etc . . .) to support their data analysis. In 2011, Yahoo's Hadoop infrastructure consisted of 42,000 nodes and hundreds of petabytes of storage.[22]

Even with the advent of an open source package built on the innovative ideas of Google and decades of advances in hardware development and high performance computing, there was one critical step remaining in the process of democratizing big data analysis – making these advancements accessible. Multiple companies sprung up around the open source Hadoop ecosystem to provide technical support to help apply these approaches to the wide range of business challenges in the marketplace. Companies like Cloudera, Hortonworks, and MapR have helped make these powerful technologies available to companies who had data but not the technological workforce to analyze them.

Collectively, these advances have moved analysis of large datasets from a

19. Jeff Dean & Sanjay Ghemawat, *MapReduce: Simplified Data Processing on Large Clusters*, COMM. OF THE ACM, Jan. 2008, at 107.

20. FAY CHANG ET AL., BIGTABLE: A DISTRIBUTED STORAGE SYSTEM FOR STRUCTURED DATA (2006).

21. HADOOP, http://hadoop.apache.org/.

22. Derrick Harris, *The history of Hadoop: From 4 nodes to the future of data*, GIGAOM (Mar. 4, 2013), from http://gigaom.com/2013/03/04/the-history-of-hadoop-from-4-nodes-to-the-future-of-data/.

high-cost, specialized industry fifteen years ago to one that is accessible and affordable to any company. The impact on the business community has been profound with the big data marketplace for hardware, software and services in 2012 topping $11 billion and projected to grow at a greater than 30% rate for the coming four years.[23]

Stepping back from the hype there a several fundamental changes that have resulted from the democratization of big data. First, the proliferation of sensors and the Internet have led to many new streams of data. There is simply more data available to be collected and analyzed now than there has been in years past. Second, the cost of storage is now often lower than the cost of deciding what to throw away. This fact combined with the belief that some data, while seemingly unimportant when it is collected, can yield great value when connected to other data in the future, will lead companies to increasingly default to storing data rather than discarding it. Finally, cloud-based big data providers provide zero capital investment ways to start analyzing big data. This provides accessible, scalable methods for storage and analysis.

Though less apparent in revenue figures, the impact on the national security community has been profound as well since these new big data analysis technologies offer new ways to address challenges of dealing with large datasets for the defense, intelligence, and homeland security realms. The national security community faces data analysis challenges in many domains that require sifting through large volumes of data to find a signal of interest. Whether it is finding evidence of money laundering in financial transaction data, suspect shipping containers in transportation manifest data or malicious cyber activity in internet log data, the government faces challenges that require storing and analyzing massive quantities of data. The democratization of big data puts these new analytic tools within reach of government agencies as well as companies and offers the ability for improved mission performance.

In other words, one area where democratization may affect national security is in terms of cost. Given our current budget constrained environment, anything that is becoming more cost effective will be welcomed with open arms. As noted above, the major news item regarding the democratization of big data is that is now much cheaper to store and analyze data. Small businesses can now afford the analytical tools, services and experts whereas before this storage and analysis was prohibitively expensive. One of the benefits, then, may be that the U.S. Government, like the small business owner, can store and analyze more data at lower cost. This is magnificent news to those who are wrestling with budgets.

Both the National Security Agency and Central Intelligence Agency have publicly cited the advent of big data analytic techniques as core to their technology roadmaps going forward. "Revolutionize Big Data Exploitation" is

23. Jeff Kelly et al., *Big Data Vendor Revenue And Market Forecast 2012–2017*, WIKIBON (Oct. 19, 2013), http://wikibon.org/wiki/v/Big_Data_Vendor_Revenue_and_Market_Forecast_2012-2017.

one of the "Four Big Bets" that the CIA is currently making in its technology strategy.[24] Similarly, NSA's Chief Information Officer, Lonny Anderson, points out that the NSA "is at the forefront of the IC's efforts to implement big data analytics and an underlying cloud computing infrastructure."[25]

One of the purposes of intelligence, particularly defense intelligence, is to inform commanders' decisions by providing insightful intelligence. DoD Joint Publication 2-0 on Joint Intelligence lays out the intelligence analysis process by which raw data is collected from the operational environment by sensors.[26] This data is, in turn, distilled through processing and exploitation into information and finally, the information is refined through analysis and production into intelligence.[27]

That analytical process is not new. The military has been using data analytics within the context of the defense intelligence cycle long before big data and cloud-computing services became democratized. Now, thanks to that exciting technological development, the defense intelligence cycle can deliver to warfighters and commanders better intelligence at lower cost. For instance, ground troops can plan mounted and dismounted patrol routes based on queries of attack types in an area of operation. Anti-piracy task forces in the Indian Ocean could, using similar analytics, plan more efficient coverage of that vast space with queries and analysis of types and times of pirate activities. Remote piloted vehicle (RPV) coverage of border areas or drug routes could be improved both domestically and abroad. In sum, big data democratization affects the battlefield, not just the budget, because technological tools are more accessible now then they were a few years ago.

In sum, advances in big data offer the national security community, and the defense department in particular, the capability to both distill and relate information more efficiently than could be done with conventional methods. Whether it is identifying objects of interest from RPV video streams or connecting disparate pieces of information that allow for the detection of potential terrorists, the national security community has high hopes for the operational benefits of big data.

---

24. Matt Sledge, *CIA's Gus Hunt On Big Data: We "Try To Collect Everything And Hang On To It Forever."* HUFFINGTON POST (Mar. 20, 2013), http://www.huffingtonpost.com/2013/03/20/cia-gus-hunt-big-data_n_2917842.html.

25. George I. Seffers, *Big Data in Demand for Intelligence Community*, AFCEA (Jan. 4, 2013), http://www.afcea.org/content/?q=node/10510.

26. JOINT CHIEFS OF STAFF, JOINT PUB. 2-0, JOINT INTELLIGENCE (2007), *available at* http://www.dtic.mil/doctrine/new_pubs/jp2_0.pdf.

27. "Raw data by itself has relatively limited utility. However, when data is collected from a sensor and processed into an intelligible form, it becomes information and gains greater utility. Information on its own is a fact or a series of facts that may be of utility to the commander, but when related to other information already known about the operational environment and considered in the light of past experience regarding an adversary, it gives rise to a new set of facts, which may be termed 'intelligence.' The relating of one set of information to another or the comparing of information against a database of knowledge already held and the drawing of conclusions by an intelligence analyst, is the foundation of the process by which intelligence is produced." *Id.* at I-2.

***