Big-Data Computing: Creating revolutionary breakthroughs in commerce, science, and society

Randal E. Bryant
Carnegie Mellon University

Randy H. Katz
University of California, Berkeley

Edward D. Lazowska
University of Washington

Version 8: December 22, 2008[1]

Motivation: Our Data-Driven World

Advances in digital sensors, communications, computation, and storage have created huge collections of data, capturing information of value to business, science, government, and society. For example, search engine companies such as Google, Yahoo!, and Microsoft have created an entirely new business by capturing the information freely available on the World Wide Web and providing it to people in useful ways. These companies collect trillions of bytes of data every day and continually add new services such as satellite images, driving directions, and image retrieval. The societal benefits of these services are immeasurable, having transformed how people find and make use of information on a daily basis.

Just as search engines have transformed how we access information, other forms of *big-data computing* can and will transform the activities of companies, scientific researchers, medical practitioners, and our nation's defense and intelligence operations. Some examples include:

- Wal-Mart recently contracted with Hewlett Packard to construct a data warehouse capable of storing 4 *petabytes* (4000 trillion bytes) of data, representing every single purchase recorded by their point-of-sale terminals (around 267 million transactions per day) at their 6000 stores worldwide. By applying *machine learning* to this data, they can detect patterns indicating the effectiveness of their pricing strategies and advertising campaigns, and better manage their inventory and supply chains.

- Many scientific disciplines have become data-driven. For example, a modern telescope is really just a very large digital camera. The proposed Large Synoptic Survey Telescope (LSST) will scan the sky from a mountaintop in Chile, recording 30 trillion bytes of image data every day – a data volume equal to *two entire Sloan Digital Sky Surveys daily*! Astronomers will apply massive computing power to this data to probe the origins of our universe. The Large Hadron Collider (LHC), a particle accelerator that will revolutionize our understanding of the workings of the Universe, will generate 60 terabytes of data per day – 15 petabytes (15 million gigabytes) annually. Similar *eScience* projects are proposed or underway in a wide variety of other disciplines, from biology to environmental science to oceanography. These projects generate such enormous data sets that automated analysis is required. Additionally, it becomes impractical to replicate copies at the sites of individual research groups, so investigators pool their resources to construct a large data center that can run the analysis programs for all of the affiliated scientists.

[1] For the most current version of this essay, as well as related essays, visit http://www.cra.org/ccc/initiatives

- Modern medicine collects huge amounts of information about patients through imaging technology (CAT scans, MRI), genetic analysis (DNA microarrays), and other forms of diagnostic equipment. By applying *data mining* to data sets for large numbers of patients, medical researchers are gaining fundamental insights into the genetic and environmental causes of diseases, and creating more effective means of diagnosis.

- Understanding the environment requires collecting and analyzing data from thousands of sensors monitoring air and water quality and meteorological conditions, another example of eScience. These measurements can then be used to guide simulations of climate and groundwater models to create reliable methods to predict the effects of long-term trends, such as increased CO_2 emissions and the use of chemical fertilizers.

- Our intelligence agencies are being overwhelmed by the vast amounts of data being collected through satellite imagery, signal intercepts, and even from publicly available sources such as the Internet and news media. Finding and evaluating possible threats from this data requires "connecting the dots" between multiple sources, e.g., to automatically match the voice in an intercepted cell phone call with one in a video posted on a terrorist website.

- The collection of all documents on the World Wide Web (several hundred trillion bytes of text) is proving to be a corpus that can be mined and processed in many different ways. For example, language translation programs can be guided by statistical language models generated by analyzing billions of documents in the source and target languages, as well as multilingual documents, such as the minutes of the United Nations. Specialized web crawlers scan for documents at different reading levels to aid English-language education for first graders to adults. A conceptual network of noun-verb associations has been constructed based on word combinations found in web documents to guide a research project at Carnegie Mellon University in which fMRI images are used to detect how human brains store information.

These are but a small sample of the ways that all facets of commerce, science, society, and national security are being transformed by the availability of large amounts of data and the means to extract new forms of understanding from this data.

Big-Data Technology: Sense, Collect, Store, and Analyze

The rising importance of big-data computing stems from advances in many different technologies:

Sensors: Digital data are being generated by many different sources, including digital imagers (telescopes, video cameras, MRI machines), chemical and biological sensors (microarrays, environmental monitors), and even the millions of individuals and organizations generating web pages.

Computer networks: Data from the many different sources can be collected into massive data sets via localized sensor networks, as well as the Internet.

Data storage: Advances in magnetic disk technology have dramatically decreased the cost of storing data. For example, a one-terabyte disk drive, holding one trillion bytes of data, costs around $100. As a reference, it is estimated that if all of the text in all of the books in the Library of Congress could be converted to digital form, it would add up to only around 20 terabytes.

Cluster computer systems: A new form of computer systems, consisting of thousands of "nodes," each having several processors and disks, connected by high-speed local-area networks, has become the chosen hardware configuration for data-intensive computing systems. These clusters provide both the storage capacity for large data sets, and the computing power to organize the data, to analyze it, and to respond to queries about the data from remote users. Compared with traditional high-performance computing (e.g., supercomputers), where the focus is on maximizing the raw computing power of a system, cluster computers are designed to maximize the reliability and efficiency with which they can manage and analyze very large data sets. The "trick" is in the software algorithms – cluster computer systems are composed of huge numbers of cheap commodity hardware parts, with scalability, reliability, and programmability achieved by new software paradigms.

Cloud computing facilities: The rise of large data centers and cluster computers has created a new business model, where businesses and individuals can *rent* storage and computing capacity, rather than making the large capital investments needed to construct and provision large-scale computer installations. For example, Amazon Web Services (AWS) provides both network-accessible storage priced by the gigabyte-month and computing cycles priced by the CPU-hour. Just as few organizations operate their own power plants, we can foresee an era where data storage and computing become utilities that are ubiquitously available.

Data analysis algorithms: The enormous volumes of data require automated or semi-automated analysis – techniques to detect patterns, identify anomalies, and extract knowledge. Again, the "trick" is in the software algorithms - new forms of computation, combining statistical analysis, optimization, and artificial intelligence, are able to construct statistical models from large collections of data and to infer how the system should respond to new data. For example, Netflix uses machine learning in its recommendation system, predicting the interests of a customer by comparing her movie viewing history to a statistical model generated from the collective viewing habits of millions of other customers.

Technology and Application Challenges

Much of the technology required for big-data computing is developing at a satisfactory rate due to market forces and technological evolution. For example, disk drive capacity is increasing and prices are dropping due to the ongoing progress of magnetic storage technology and the large economies of scale provided by both personal computers and large data centers. Other aspects require more focused attention, including:

High-speed networking: Although one terabyte can be stored on disk for just $100, transferring that much data requires an hour or more within a cluster and roughly a day over a typical “high-speed” Internet connection. (Curiously, the most practical method for transferring bulk data from one site to another is to ship a disk drive via Federal Express.) These bandwidth limitations increase the challenge of making efficient use of the computing and storage resources in a cluster. They also limit the ability to link geographically dispersed clusters and to transfer data between a cluster and an end user. This disparity between the amount of data that is practical to store, vs. the amount that is practical to communicate will continue to increase. We need a “Moore’s Law” technology for networking, where declining costs for networking infrastructure combine with increasing bandwidth.

Cluster computer programming: Programming large-scale, distributed computer systems is a longstanding challenge that becomes essential to process very large data sets in reasonable amounts of time. The software must distribute the data and computation across the nodes in a cluster, and detect and remediate the inevitable hardware and software errors that occur in systems of this scale. Major innovations have been made in methods to organize and program such systems, including the MapReduce programming framework introduced by Google. Much more powerful and general techniques must be developed to fully realize the power of big-data computing across multiple domains.

Extending the reach of cloud computing: Although Amazon is making good money with AWS, technological limitations, especially communication bandwidth, make AWS unsuitable for tasks that require extensive computation over large amounts of data. In addition, the bandwidth limitations of getting data in and out of a cloud facility incur considerable time and expense. In an ideal world, the cloud systems should be geographically dispersed to reduce their vulnerability due to earthquakes and other catastrophes. But, this requires much greater levels of interoperability and data mobility. The OpenCirrus project is pointed in this direction, setting up an international testbed to allow experiments on interlinked cluster systems. On the administrative side, organizations must adjust to a new costing model. For example, government contracts to universities do not charge overhead for capital costs (e.g., buying a large machine) but they do for operating costs (e.g., renting from AWS). Over time, we can envision an entire ecology of cloud facilities, some providing generic computing capabilities and others targeted toward specific services or holding specialized data sets.

Machine learning and other data analysis techniques: As a scientific discipline, machine learning is still in its early stages of development. Many algorithms do not scale beyond data sets of a few million elements or cannot tolerate the statistical noise and gaps found in real-world data. Further research is required to develop algorithms that apply in real-world situations and on data sets of trillions of elements. The automated or semi-automated analysis of enormous volumes of data lies at the heart of big-data computing for all application domains.

Widespread deployment: Until recently, the main innovators in this domain have been companies with Internet-enabled businesses, such as search engines, online retailers, and social networking sites. Only now are technologists in other organizations (including universities) becoming familiar with the capabilities and tools. Although many organizations are collecting large amounts of data, only a handful are making full use of the insights that this data can provide. We expect "big-data science" – often referred to as eScience – to be pervasive, with far broader reach and impact even than previous-generation computational science.

Security and privacy: Data sets consisting of so much, possibly sensitive data, and the tools to extract and make use of this information give rise to many possibilities for unauthorized access and use. Much of our preservation of privacy in society relies on current inefficiencies. For example, people are monitored by video cameras in many locations – ATMs, convenience stores, airport security lines, and urban intersections. Once these sources are networked together, and sophisticated computing technology makes it possible to correlate and analyze these data streams, the prospect for abuse becomes significant. In addition, cloud facilities become a cost-effective platform for malicious agents, e.g., to launch a botnet or to apply massive parallelism to break a cryptosystem. Along with developing this technology to enable useful capabilities, we must create safeguards to prevent abuse.

Leadership

This is an area where industry has been in the lead, especially the Internet-enabled service companies. These companies are investing billions of dollars in computing infrastructure that dwarf the world's largest traditional supercomputing installations. The main innovators in the configuration and programming of cluster computing systems have been at Google, Yahoo!, and Amazon. Other companies, from retailers to financial services, are taking notice of the business advantages and the operating efficiencies these companies are finding.

University researchers have been relatively late to this game, due to a combination of lack of access to large-scale cluster computing facilities and to a lack of appreciation for the new insights that can be gained by scaling up to terabyte-scale data sets. This situation is rapidly changing through access to facilities and training, and due to the successes of their research counterparts in industry. Google, IBM, Yahoo!, and Amazon have provided access to some of their computing resources for students and researchers, using a cloud computing model. This has been enough to whet appetites but not nearly enough to satisfy the potential needs for widespread application of data-intensive computing. Many large-scale scientific projects are formulating plans for how they will manage and provide computing capacity for their collected data. Spirited debate is taking place between proponents of this new approach to data management and computing, with those pursuing more traditional approaches such as database technology and supercomputing. A lack of sufficient research funding is the major obstacle to getting greater involvement among university researchers.

Unfortunately, leadership from government agencies in this area has been mixed at best. The NSF has embraced data-intensive computing enthusiastically, with several new funding initiatives both within the CISE Directorate and with ties to other fields of science. However, their recent budgets have been so constrained that these programs have not been able to scale up to their full potential. The Cyber-enabled Discovery and Innovation program is scheduled to receive $26M for FY09. This is a step in the right direction, but further funding growth is required. Both the DoD and the DoE *should* be deeply involved in the development and deployment of big-data computing, since it will be of direct benefit to many of their missions. Sadly, neither of these agencies has been a driving force for innovations in computing technology in recent times. Both are making heavy investments in traditional high-performance computing infrastructure and approaches, but very little in new eScience facilities and technologies.

Recommendations

Investments in big-data computing will have extraordinary near-term and long-term benefits. The technology has already been proven in some industry sectors; the challenge is to extend the technology and to apply it more widely. Below we list specific actions by the federal government that would greatly accelerate progress.

Immediate Actions

Specific funding over the next two years could greatly stimulate the development, deployment, and application of big-data computing. In making these short-term expenditures, we must take care to make investments that lead to substantial short *and* long-term benefits. In particular, computer hardware is fundamentally a depreciating asset, with systems having a useful lifetime of 3-5 years. So, investing heavily in machines in anticipation of a future computing need is wasteful. We should spend only as much as can

be beneficially used right away. Networking infrastructure has a longer-term useful lifetime (8-10 years), and so these investments can have longer term value. Finally, knowledgeable people are an appreciating asset. Providing funding opportunities for more students and researchers to work with big-data computing will yield very high returns over many years.

Below we list some proposals for near-term initiatives. A much more careful analysis would be required to determine an optimal set of priorities with firm budgets.

- Invest in higher capacity networking infrastructure, both in the backbones as well as access by major research universities and government labs. Assist cloud service providers in connecting their systems to this high capacity network backbone (e.g., with tax incentives). Total cost: ~$100M.

- Invest in upgrades in high-performance computing sites, such as national laboratories and supercomputer centers, with a specific requirement that the additional funds be directed toward better serving the data-intensive storage and computing needs of their users. This could include better storage servers, as well as cluster computing systems closely linked to the high-performance machines to aid in analyzing the data being generated. Total cost: ~$100M.

- Establish a networked collection of 10 cluster systems, in the spirit of the OpenCirrus initiative, but with greater coordination, oversight, and network bandwidth between them. These systems should be geographically dispersed, e.g., at multiple universities and research laboratories. Some of these machines would be made available to researchers pursuing applications of big-data computing, while the others would be made available to systems researchers, exploring both the operations of individual clusters, as well as connecting multiple clusters into a cloud. These researchers need access to machines where they can control, configure, and frequently crash the low-level system software. Opportunities should be created for industry partners to collaborate by connecting their own clusters to the network. Cost for hardware: around $50M. Operating cost: around $10M/year. Total cost: ~$100M.

- Increase the NSF budget enough to enable their initiatives in data-intensive computing to be fully realized. These include the Cyber-Enabled Discovery and Innovation program, which connects computing disciplines with other areas of science (an excellent source of eScience programs), the Computing Expeditions program, and several others. These programs all received a very large number of high quality proposals that could not be funded due to tight budgets, and so an injection of funds could allow many projects to proceed in a short amount of time. Total cost: ~$75M.

Longer Term Actions

Specific actions that the federal government could take include:

- Give the NSF a large enough budget increase that they can foster efforts in big-data computing without having to cut back on other programs. Research thrusts within computing must cover a wide range of topics, including: hardware and system software design; data-parallel programming and algorithms; automatic tuning, diagnosis and repair in the presence of faults; scalable machine learning algorithms; security and privacy; and applications such as language translation and computer vision. Interdisciplinary programs should marry technologists with applications experts with access to extremely large datasets in other fields of science, medicine, and engineering.

- Reconsider the current plans to construct special-purpose data centers for the major eScience programs. Possible economies of scale could be realized by consolidating these into a small number of "super data centers" provisioned as cloud computing facilities. This approach would provide opportunities for technologists to interact with and support domain scientists more effectively. These efforts should be coupled with large-scale networking research projects.

- Renew the role of DARPA in driving innovations in computing technology and its applications. Projects could include applications of interest to the DoD, including language understanding, image and video analysis, and sensor networks. In addition, they should be driving the fundamental technology required to address problems at the scale faced by the DoD. Both the systems and data analysis technologies are clearly "dual use."

- Sensitize the DoD to the potential for technological surprise. An adversary with very modest financial resources could have access to supercomputer-class computer facilities. $100 buys 1000 processors for 1 hour on AWS. $100M – considerably less than the cost of single modern strike fighter – buys one billion processor-hours. Cloud computing must be considered a strategic resource, and it is essential that the US stays in the lead in the evolution and application of this technology.

- Get the DoE to look beyond traditional high-performance computing in carrying out their energy and nuclear weapons missions. Many of their needs could be addressed better and more cost effectively by cluster computing systems, possibly making use of cloud facilities.

- Encourage the deployment and application of big-data computing in all facets of the government, ranging from the IRS (tax fraud detection), intelligence agencies (multimedia information fusion), the CDC (temporal and geographic tracking of disease outbreaks), and the Census Bureau (population trends).

- Make fundamental investments in our networking infrastructure to provide ubiquitous, broadband access to end users and to cloud facilities.

Big-data computing is perhaps the biggest innovation in computing in the last decade. We have only begun to see its potential to collect, organize, and process data in all walks of life. A modest investment by the federal government could greatly accelerate its development and deployment.