U.S. HOUSE OF REPRESENTATIVES
COMMITTEE ON SCIENCE, SPACE, AND TECHNOLOGY
SUBCOMMITTEES ON RESEARCH AND TECHNOLOGY
HEARING CHARTER

Next Generation Computing and Big Data Analytics

Wednesday, April 24, 2013
10:00 a.m. – 12:00 p.m.
2318 Rayburn House Office Building

Purpose

On Wednesday, April 24, 2013, the House Committee on Science, Space, and Technology's Research and Technology Subcommittees will examine how advancements in information technology and data analytics enable private and public sector organizations to utilize mass volumes of data to provide greater value to their customers and citizens, spurring new product and service innovations. The hearing will focus on innovative data analytics capabilities, research and development efforts, management challenges, and workforce development issues associated with the "Big Data" phenomenon.

Witnesses

- **Dr. David McQueeney,** Vice President, Technical Strategy and Worldwide Operations, IBM Research
- **Dr. Michael Rappa,** Executive Director of the Institute for Advanced Analytics, Distinguished University Professor, North Carolina State University
- **Dr. Farnam Jahanian,** Assistant Director for the Computer and Information Science and Engineering (CISE) Directorate, National Science Foundation (NSF)

Overview

Unprecedented volumes of complex and diverse data sets are being generated daily across a range of industries and public sector organizations. The term "Big Data" encompasses the challenge of collecting, analyzing and disseminating the massive data sets that are currently being generated and stored. Private industry and government officials are seeking ways to harness, analyze, and exploit these data sets in ways that provide greater value to their customers and citizens. While Big Data is a relatively new term, the problem is not. What is changing is both the volume of the data and the pressure to find technological solutions to managing, storing, and utilizing that data.

The McKinsey Global Institute estimated that global enterprises stored more than seven exabytes of new data on disk drives in 2010, and that consumers stored more than six exabytes on personal computers and laptops.[1] An Exabyte is 10^{18} bytes or one billion gigabytes. As a frame

[1] *Big data: The next frontier for innovation, competition, and productivity*, McKinsey Global Institute, May 2011, McKinsey and Company.

of reference, one exabyte of data is more than 4,000 times the amount of data stored in the U.S. Library of Congress.

Given the evolution of computing power and analytical capabilities, overcoming the challenge of data management presents significant need for technological innovation. High performance computing can process mass, complex sets of data at a greater rate; mathematicians and statisticians are developing new algorithms to analyze data; and data analytics professionals are employing new techniques to extract value from data.

Big Data has profound implications for a range of industries. For example, health care data can enable care providers to monitor health trends and evaluate different treatments, energy data can inform power distribution creating greater efficiencies, transportation data can be used to mitigate traffic congestion, and information technology data can identify potential cyber threats.

In addition, technological advances allow scientists to both collect and analyze data at a significantly faster rate. Examples include advancements in human genome sequencing, digital astronomy data, and particle physics.

Industry and Big Data

Big Data represents a significant growth area for private industry. In recent years, industry spending on data analytics and management has increased approximately 10 percent a year.[2]

Companies utilize data analytics to manage supply chains, target marketing based on user preferences, provide airline fare prediction services for consumers, and reduce costs by identifying operating inefficiencies, among a multitude of other uses.

Information Communications and Technology (ICT) companies and management consulting companies are providing a range of Big Data capabilities, including software, hardware, and analytics services. Industry customers of Big Data products and services, including health care, transportation, agriculture, and retail companies are identifying ways to increase yields, cut costs, and increase customer retention. ICT companies also work closely with government and academia to build high performance computers and software systems, which enable cutting edge Big Data scientific research and development initiatives.

Big Data Workforce Development

McKinsey has projected the United States will need an additional 140,000 to 190,000 professionals with significant technical depth in data analytics, and the need for an additional 1.5 million managers and analysts who can work effectively with big data analysis by 2018.[3]

To address anticipated workforce demands, colleges and universities are recognizing the value in providing students with education and training in Big Data-related disciplines. Such programs provide instruction in a broad spectrum of Big Data-related disciplines including data management, mathematical and statistical methods for data modeling, and techniques for data

[2] "*A special report on managing information: Data, data everywhere*," The Economist; February 25, 2010.
[3] *Big data: The next frontier for innovation, competition, and productivity*, McKinsey Global Institute, May 2011, McKinsey and Company.

visualization in support of business decision making. Although some institutions have initiated these types of degree programs, overall they are still relatively rare.

Federal Big Data Research and Development Initiatives

On March 29, 2012, the Obama Administration unveiled its "Big Data Research and Development Initiative,"[4] announcing more than $200 million in new funding to improve the tools and techniques needed to access, organize, and glean discoveries from huge volumes of digital data.

Six federal departments and agencies, including the National Science Foundation (NSF), National Institutes of Health (NIH), the Department of Defense (DOD) and Defense Advanced Research Projects Agency (DARPA), Department of Energy (DOE), and the U.S. Geological Survey (USGS) are participating in this initiative.

National Science Foundation and Big Data

The NSF Computer and Information Sciences and Engineering Directorate (CISE) supports investigator-initiated research in all areas of computer and information science and engineering, helps develop and maintain cutting-edge national computing and information infrastructure for research and education generally, and contributes to the education and training of the next generation of computer scientists and engineers.

CISE supports Big Data investments in foundational research, cyberinfrastructure, education and workforce development needs, and in efforts to support interdisciplinary research.

Core Techniques and Technologies for Advancing Big Data Science & Engineering.

As part of the President's Big Data Initiative, the NSF/NIH Core Techniques and Technologies for Advancing Big Data Science & Engineering (BIGDATA) is offering research grants to accelerate the progress of scientific discovery and innovation; lead to new fields of inquiry that would not otherwise be possible; encourage the development of new data analytic tools and algorithms; facilitate scalable, accessible, and sustainable data infrastructure; increase understanding of human and social processes and interactions; and promote economic growth and improved health and quality of life.

Yellowstone, Blue Waters, Gordon, and Stampede Supercomputers

NSF also advances Big Data computational research and development through the Yellowstone, Blue Waters, and Stampede Supercomputers, in partnership with the University of Wyoming, the University of Illinois, the University of California, San Diego, and the University of Texas at Austin, respectively.

Yellowstone is the petascale computing resource in the National Center for Atmospheric Research (NCAR)-Wyoming Supercomputing Center (NWSC), which opened in October 2012.

[4] http://www.whitehouse.gov/blog/2012/03/29/big-data-big-deal

(A petascale refers to a computer system capable of reaching performance in excess of one petaflops, i.e. one quadrillion floating point operations per second.) The NWSC provides advanced computing services to scientists studying a broad range of disciplines, including weather, climate, oceanography, air pollution, space weather, computational science, energy production, and carbon sequestration.

The Blue Waters supercomputer provides sustained performance of one petaflop on a range of real-world science and engineering applications. Blue Waters enables scientists and engineers across the country to tackle a wide range of challenging problems, from predicting the behavior of complex biological systems to simulating the evolution of the cosmos.

The Gordon Compute Cluster is a unique data-intensive supercomputer sponsored by NSF, went into production January 1, 2012. Large graph problems, data mining, genome assembly, database applications, and quantum chemistry are some of the fields of research benefitting from Gordon's unique architecture.

Stampede was officially dedicated in March 2013 at the University of Texas at Austin's Advanced Computing Center (TACC). Stampede will have a peak performance of 10 petaflops. Research programs already being conducted at Stampede include seismic hazard mapping, ice sheet modeling, improving the imaging quality of brain tumors, and carbon dioxide capture and conversion.

Department of Energy Scalable Data Management, Analysis and Visualization (SDAV) Institute

On March 29, 2012, the DOE announced $25 million to establish the Scalable Data Management, Analysis and Visualization (SDAV) Institute to extract knowledge and insights from large and complex collections of digital data. Led by the Energy Department's Lawrence Berkeley National Laboratory, the SDAV Institute brings together the expertise of six national laboratories and seven universities to develop new tools to help scientists manage and visualize data on the Department's supercomputers, which will further streamline the processes that lead to discoveries made by scientists using the Department's research facilities.[5]

The SDAV Institute helps scientists extract insights from today's increasingly massive research datasets by assisting researchers in using state-of-the art software tools for data analysis on these supercomputers – ranging from superfast search engines to sophisticated visualization software that enables researchers to literally picture and "see" complex relations among data points. The Energy Department supports some of the world's fastest supercomputers located at Argonne, Oak Ridge, and Lawrence Berkeley National Laboratories, which are used by scientists from a wide range of fields.

National Institute of Standards and Technology (NIST) Big Data Initiatives

The NIST Information Technology Laboratory conducts a number of activities related to Big Data through its Computer Security Division. NIST conducts research on the science behind Big

[5] http://www.sdav-scidac.org/report.html

Data, measurement tools to advance Big Data, privacy, and the security of Big Data infrastructure. This effort includes convening industry and interested stakeholders together to explore the challenges of Big Data, including common terms and taxonomies for use by the field, and identification of areas where research is needed.

Specifically, NIST has worked in areas of convergence between cloud computing and data, particularly on the interoperability of cloud platforms. Although NIST does not play a visible role in the Administration's Big Data Research and Development Initiative, it supports the creation of many of the analytical tools to address the challenges of Big Data in both the public and private sectors.

Areas for Examination

Witnesses have been asked to describe private and public Big Data research and development efforts; applications of Big Data initiatives; and management challenges, including workforce development issues.

In addition, the Committee will seek to determine: how federal big data research projects are coordinated across participating agencies; how the public and private sectors manage privacy concerns as part of Big Data initiatives; how federal privacy laws, such as health privacy laws, affect opportunities to gain information from data; and how Congress should prioritize Big Data research initiatives in federal research budgets.