Government Business Council
Briefings

A Briefing
from GBC's
Research Analysts
July 2012

5 of the Most Innovative Government Big Data Projects

Anyone who has ever asked Siri, Apple's automated personal assistant, for directions or to make an appointment has used a government big data project. Siri, just like the computer, internet, cell phone, and microwave before it, has its origins in a Pentagon laboratory. After being spun off as a Silicon Valley start-up, Siri was eventually purchased by Apple in 2010. Siri's claim to fame is its ability to access and analyze massive sums of data. And the more data Siri is exposed to, the smarter it becomes.[1]

> "According to one industry estimate, 90 percent of all data in the world has been created in the past two years... To deal with this massive growth, the federal government recently announced investments in big data projects totaling $200 million."

Siri's debut has coincided with the largest explosion of data in history. According to one industry estimate, 90 percent of all data in the world has been created in the past two years.[2] Another estimate suggests that every year, the amount of data in the world grows by 40 percent.[3] To deal with this massive growth, the federal government recently announced investments in big data projects totaling $200 million.[4]

The big data projects are diverse and numerous, spanning many departments and numbering well over 80 unique projects. They range from sharing measurements at the Large Hadron Collider with scientists around the world to improving machine reading and learning, Siri's forte. Some projects, however, are set to truly push the envelope and make full use of recent advances in big data computation and collection.

These are five of the most of innovative and impactful projects the government is investing in, and how they aim to drive further innovation.

1. 1000 Genomes Project

The 1000 Genomes Project began in 2008 as an international effort to catalog all three billion DNA bases in the human genome. Seventy-five companies and organizations have collected over 200 terabytes of data on more than 2,500 individuals.[5]

This information is invaluable to thousands of researchers all over the world. However, as Lisa. D. Brooks, Program Director for the Genetic Variation Program at the Nation Institutes of Health (NIH), notes, "Previously, researchers wanting access to public data sets such as the 1000 Genomes Project had to download them from government data centers to their own systems, or have the data physically shipped to them on discs. This process took a long time, and that's assuming a lab had the bandwidth to download the data and sufficient storage and compute infrastructure to hold and analyze the data once they had it."[6]

In order to circumvent this problem, NIH contracted with Amazon to create the Amazon web service (AWS). This move makes the data, which is enough to fill about 30,000 DVDs, easily

and quickly accessible around the world.[7] Simple access to this data is expected to spur new advances in the study of the human genome and potential new medical applications.

> "Gentry showed it was theoretically possible to compute fully encrypted data, where only the owner was able to decrypt the data with an encryption key."

2. Programming Computation on Encrypted Data (PROCEED)

The growth of cloud computing has been perhaps the most dominant tech trend over the past few years. Yet, wherever there is cloud computing, there are critics barking about its poor security and vulnerability to attack. The Defense Advanced Research Projects Agency (DARPA), though, is investing in ways to make massive amounts of information computed on the cloud secure, removing fears of hacking and interception of data at the point of decryption.

One of the most promising strategies is fully homomorphic encryption (FHE), which allows third parties to compute the data without decrypting it. If this term sounds familiar, it is perhaps because an IBM researcher, Craig Gentry, made headlines in 2009 by discovering a method for computers to process data without decrypting it. Gentry showed it was theoretically possible to compute fully encrypted data, where only the owner was able to decrypt the data with an encryption key.[8]

However, computation of encrypted data is incredibly slow, nearly 10 orders of magnitude slower than computing decrypted data.[9] One of the most important consequences of PROCEED, if successful, will be the decreased computation time of encrypted data. This would allow agencies to process massive data sets on the cloud completely and securely, and remove the major barrier hindering full transition to cloud computing.

> "[BioSense] uses symptomatic data gathered from all over the country to track health issues as they evolve."

3. Bio Sense 2.0

BioSense began in 2003 as an attempt to create an "integrated national public health surveillance system for early detection and rapid assessment of potential bioterrorism-related illness."[10] Originally mandated in the Public Health Security and Bioterrorism Preparedness and Response Act of 2002, BioSense is now being expanded and updated by the Center for Disease Control, its parent organization, to cover all manner of public health tracking issues at the state, local, and national level.

BioSense 2.0 is a collaborative system between all levels of government hosted on the cloud that provides data simply and readily to end-users. It uses symptomatic data gathered from all over the country to track health issues as they evolve. The system is also designed to allow users full access even when using query software, such as R, a free statistical computing environment.[11] It is receiving accolades from many public health scholars, such as John Halamka, CIO of Harvard Medical School.[12] In addition, all the data stored on the cloud is in complete compliance with the Federal Information Security Management Act (FISMA), removing many concerns of privacy activists.[13]

4. Genomic Information System for Integrated Science (GenISIS) and the Million Veterans Program (MVP)

GenISIS and the Million Veterans Program are projects supported by the Department of Veterans' Affairs to make greater use of veteran genomic information and ultimately improve patient care. Having full genomic information on veterans will give doctors far greater information when treating patients. The projects will also make vast amounts of genomic data available for

> Researchers will have access to the genomic data, in addition to clinical data, research data, biological data, and medical records.

secondary research and analysis, potentially leading to new or improved treatments. "Thus, the short-term goal for GenISIS is to create and support a knowledge base that would facilitate independent research projects and collaborative repurposing of data. The vision for GenISIS for the longer term is focused on patient care, integrating clinical care and research activities for improved patient outcomes."[14]

In order to provide information for such a database, the Million Veterans Program recruits volunteer veterans to contribute blood samples. These samples are processed for genotyping and genetic sequencing, and eventually added to the veteran's "phenotype" in their health records. Researchers will also have access to the genomic data, in addition to clinical data, research data, biological data, and medical records. These records are traditionally housed in separate compartments, but now researchers will have access to all records from one source.[15]

The effects of such a project will be to move medical research past targeted hypothesis testing, and into a larger collaborative research effort. Data will be easily available from experiments, meaning that researchers will often not have to reinvent the wheel, so to speak.[16] Ultimately, these projects should help to drastically improve research efforts and the patient care veterans and active-duty soldiers receive.

> The Virtual Laboratory Environment seeks to give FDA employees new mobile capabilities, and allow them to "basically make any location a virtual laboratory with advanced capabilities in a matter of hours.
>
> - Big Data Fact Sheet, Office of Science and Technology Policy

5. Virtual Laboratory Environment (VLE)

The Food and Drug Administration (FDA) has a unique challenge in that, many times, its work sites are highly variable, or not in the comfort of a modern office. FDA officials work sites range from farmland to urban dockyards. Working in such locations poses many problems, one of which is the challenge of accessing laboratory data and capabilities. However, the Virtual Laboratory Environment seeks to give FDA employees new mobile capabilities, and allow them to "basically make any location a virtual laboratory with advanced capabilities in a matter of hours."[17]

Such a laboratory will combine a virtual data network, advanced analytical and statistical tools, document management support, and tele-presence capabilities. In addition, the virtual laboratory will crowd source analytics to "predict and promote public health."[18] This environment will help the FDA complete its mission more effectively and efficiently from wherever their work takes them.

In 1969, the Apollo 11 Guidance Computer took three American astronauts to the moon with less computing power than that found in a modern cellphone. The government is now dealing with petabytes of data and discovering new ways to compute and make use of such data. Apollo 11 pushed the envelope on scientific discovery, and today, government big data projects are continuing to spur new innovations and big steps for mankind.

—By Clement Christensen with Dana Grinshpan (editor)

About NetApp

NetApp creates innovative storage and data management solutions that deliver outstanding cost efficiency and accelerate business breakthroughs. Discover our passion for helping companies around the world go further, faster at www.netapp.com.

About GBC: Briefings

As Government Executive Media Group's research division, Government Business Council Briefings are dedicated to advancing the business of government through insight and analytical independence. The GBC Briefings team conducts primary and secondary research to learn and share best practices among top government decision-makers in the tradition of *Government Executive*'s over forty years of editorial excellence.

For more information, contact Bryan Klopack, Executive Director of Research and Analysis, Government Executive Media Group, at bklopack@govexec.com.

Sources:

1. Steve Lohr, "The Age of Big Data," The New York Times, February 11, 2012, http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html?_r=1&pagewanted=all.
2. "What is Big Data?" IBM, Accessed June 28, 2012, http://www-01.ibm.com/software/data/bigdata/.
3. Kapil Bakshi, "A Primer on Big Data in State and Local Government," Government Technology, April 30, 2012, http://www.govtech.com/e-government/Primer-Big-Data-State-Local-Government.html.
4. "Fact Sheet: Big Data Across the Federal Government," Office of Science and Technology Policy, March 29, 2012, http://www.whitehouse.gov/administration/eop/ostp.
5. Sarah Perez, "Amazon and the NIH Team Up to Put Human Genome in the Cloud," TechCrunch, March 29, 2012, http://techcrunch.com/2012/03/29/amazon-and-the-nih-team-up-to-put-human-genome-in-the-cloud/.
6. "Amazon Web Services and the US National Institutes of Health Announce the Largest Catalog of Human Genetics is Now Available in the Cloud," BusinessWire, March 29, 2012, http://www.businesswire.com/news/home/20120329005721/en/Amazon-Web-Services-National-Institutes-Health-Announce.
7. Francie Diep, "World Repository of Human Genetics Will Move to Amazon's Cloud," Scientific American, April 2, 2012, http://www.scientificamerican.com/article.cfm?id=world-repository-of-human-genetics-will-move-to-amazons-cloud.
8. Andy Greenberg, "IBM's Blindfolded Computer," Forbes, July 13, 2009, http://www.forbes.com/forbes/2009/0713/breakthroughs-privacy-super-secret-encryption.html.
9. "Programming Computation on Encrypted Data (PROCEED)," Defense Advanced Research Projects Agency, Accessed June 28, 2012, http://www.darpa.mil/Our_Work/I2O/Programs/PROgramming_Computation_on_EncryptEd_Data_(PROCEED).aspx.
10. "BioSense," Centers for Disease Control and Prevention, Accessed June 28, 2012, http://www.cdc.gov/biosense/.
11. Matthew Dublin, "CDC Makes a Case for Amazon's Cloud," GenomeWeb, April 11, 2012, http://www.genomeweb.com/blog/cdc-makes-case-amazons-cloud.
12. John Halamka, "Cool Technology of the Week," Life as a Healthcare CIO, personal blog, April 13, 2012, http://geekdoctor.blogspot.com/2012/04/cool-technology-of-week.html.
13. Matthew Dublin, "CDC Makes a Case for Amazon's Cloud."
14. Sumitra Muralidhar, Office of Research and Development, Veterans Health Administration, "Veterans Health Administration: Infrastructure Development," National Academy of Sciences, 2009, http://www.ncbi.nlm.nih.gov/books/NBK32311/.
15. Ibid.
16. Ibid.
17. "Fact Sheet: Big Data Across the Federal Government."
18. Ibid.
19. Craig Nelson, "Ten Things You Didn't Know about the Apollo 11 Moon Landing," PopSci, July 13, 2009, http://www.popsci.com/military-aviation-amp-space/article/2009-06/40-years-later-ten-things-you-didnt-know-about-apollo-ii-moon-landing/.