Science Data Management: Maximizing the Yield

> Big Data is characterized not only by the enormous volume or the velocity of its generation, but also by the heterogeneity, diversity, and complexity of the data.
>
> —Suzi Iacono, US Interagency Big Data Senior Steering Group

1521-9615/13/$31.00 © 2013 IEEE
Copublished by the IEEE CS and the AIP

Judith Bayard Cushing
The Evergreen State College
James French
Corporation for National Research Initiatives

While 2012 saw *Big Data* entering American business and popular culture,[1,2] the phenomenon was definitively not news to the scientific community nor to *CiSE* readers. For years scientists have grappled with an exponential growth in data acquisition and generation, and 18 months ago the *CiSE* Big Data issue emphasized that although Big Data was creating "an extremely exciting time for scientific discovery," many challenges remained before scientists could make optimal use of data-intensive scientific computing.[3] Guest editors Francis J. Alexander, Adolfy Hoisie, and Alexander Szalay concluded that the sheer

scale of massive datasets precluded them from being easily moved about for analysis, and the heterogeneous, idiosyncratic, documented or not, structured or unstructured, data types encountered were so prevalent that "even computational scientists now agree that simply faster disk space and more and faster CPU cycles will not solve [Big Data] problems."[3]

These Big Data challenges haven't diminished since 2011. Indeed, in the US, the National Science Foundation[4,5] and the computer science research community have been aware of the many and significant challenges and are responding, and similar initiatives are underway internationally. This issue of *CiSE* features articles by five leading research teams whose scientific data-management projects respond to the challenges outlined in the *CiSE* Big Data issue and elsewhere. As we chose from among the many exciting researchers who regularly present their work in the annual Scientific and Statistical Database Management Conference (http://ssdbm.org), we settled on computer scientists who work particularly closely with domain researchers from across the natural and physical sciences on what we consider the root of the Big Science Data challenge: the data. As said so well in *Raw Data Is an Oxymoron*, data is anything but "raw" and we should "think of it as … a cultural resource … to be generated, protected, and interpreted."[6] Indeed, a focus on generating, protecting, and interpreting is a precursor to maximizing the yield of data collected by diverse science and engineering activities.

In this special issue on Scientific Data Management, Tamás Budavári and his colleagues describe SkyQuery, a system that enables astronomers to take advantage of the massive data provided by numerous telescopes. The system, which has its antecedents in the Sloan Digital Sky Survey (made possible through the efforts of Jim Gray, one of the foremost computer science researchers of the last 20 years), is helping produce a paradigm shift in astronomy. The article describes the effort to build a scalable query engine that dynamically federates the largest all-sky catalogs in parallel on a cluster of relational databases.

In "Data Near Here: Bringing Relevant Data Closer to Scientists," V.M. Megler and David Maier discuss the difficulty of knowing where and how to find and access relevant data in large scientific repositories. They call for an improvement in the tools used to archive and find such data, because unless scientists can easily access the information, large scientific repositories increasingly run the risk of losing value as their holdings expand. The authors go on to describe their novel information retrieval research and its implementation for a major oceanography project, but make the case that the approach is widely applicable to (and needed in) other scientific domains.

The next three articles—"Data Vaults: Database Technology for Scientific File Repositories" (by Milena Ivanova and her colleagues), "Collaborative Science Workflows in SQL" (by Bill Howe and his colleagues), and "SciDB: A Database Management System for Applications with Complex Analytics" (by Michael Stonebraker and his colleagues)—present three enhanced database technologies now available to scientists as alternatives to flat files, spreadsheets, and even generalized SQL database management systems (DBMS). Ivanova and Howe present their work in terms of use cases or case studies closely tied to particular domain sciences (seismology/remote sensing and observational biological oceanography, respectively). In addition to these case studies, Ivanova offers a cogent explanation of the data-management alternatives available for scientists working on relatively large projects with many collaborators; Howe's "virtual" alternative to placing data in a physical SQL database repository, on the other hand, might appeal particularly to scientists working on smaller projects. Stonebraker describes SciDB, a very large scientific database project among many computer scientists. SciDB responds to functional requirements for a DBMS that aims to alleviate the need for large science projects to "roll their own" database systems.

As guest editors of this issue, we first thank the contributing authors not only for the effort they put into preparing these articles specifically for *CiSE*, but for their dedication to helping scientists realize the promise of Big Data science and engineering. We also thank the heretofore anonymous reviewers who provided helpful feedback to both authors and editors as we compiled this issue: Shawn Bowers, Gonzaga University; James Frew, University of California–Santa Barbara; Carole Goble, University of Manchester; Richard Hooper, Consortium of Universities for the Advancement of Hydrologic Science (CUAHSI); Eduard Hovy, Carnegie Mellon University; Rebecca Koskela, DataONE at the University of New Mexico; Christine Laney, University of Texas-El Paso; Peter McCartney, National Science Foundation; Jim Myers, Rensselaer Polytechnic Institute;

Margaret O'Brien, University of California–Santa Barbara; Frank Olken, Arlington, Virginia; Eric Schulman, Institute for Defense Analyses; Mark Servilla, Long-Term Ecological Research (LTER) Network Office at the University of New Mexico; Robert Tawa, US National Ecological Observatory Network (NEON); Nancy Wiegand, University of Wisconsin-Madison; and Bruce Wilson, Oak Ridge National Laboratory.

We look forward to your comments on this issue! CiSE

References

1. V. Mayer-Schonberger and K. Cukier, *Big Data: A Revolution That Will Transform How We Live, Work, and Think*, Eamon Dolan/Houghton Mifflin Harcourt, 2013.
2. P.C. Zikopoulos et al., *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*, McGraw Hill, 2012.
3. F.J. Alexander, H. Adolfy, and A. Szalay, "Big Data," *Computing in Science & Eng.*, vol. 13, no. 6, 2011, pp. 10–13.
4. US National Science Foundation (NSF), "NSF Leads Federal Efforts in Big Data," press release 12-060, 29 Mar. 2012; www.nsf.gov/news/news_summ.jsp?cntn_id=123607.
5. US NSF, "NSF Announces Interagency Progress on Administration's Big Data Initiative," press release 12-187, 3 Oct. 2012; www.nsf.gov/news/news_summ.jsp?cntn_id=125610.
6. L. Gitelman, ed., *Raw Data Is an Oxymoron*, MIT Press, 2013.

***Judith Bayard Cushing**, a member of the CiSE editorial board, teaches computer science at The Evergreen State College in Olympia, Washington. Her research focuses on science data management across many disciplines, and she's currently the principle investigator for a US NSF-supported project called Visualization of Terrestrial and Aquatic Systems (http://blogs.evergreen.edu/vistas). Cushing has a PhD in computer science and engineering from the Oregon Graduate Institute. Contact her at judyc@evergreen.edu.*

***James French** is a computer science researcher at the Corporation for National Research Initiatives, a not-for-profit organization formed to undertake, foster, and promote research in the public interest. His focus is on strategic development of network-based information technologies. French has a PhD in computer science from the University of Virginia. Contact him at jfrench@cnri.reston.va.us.*