Extreme Data

Manish Parashar | Rutgers University
George K. Thiruvathukal | Loyola Unversity Chicago

The current era of extreme data culled from a range of diverse data sources, ranging from extreme-scale simulations to instruments, experiments, and pervasive sensors and systems (the so-called *Internet of Things*), has the potential for revolutionizing science, engineering, and society in general. Unprecedented instrumentation and the exponential growth of digital data sources, coupled with unprecedented advances in computing technologies, have the potential for fundamentally transforming our ability to understand and manage our lives and our environment.

We can envision data-driven and information-rich pervasive computational ecosystems that seamlessly, symbiotically, and opportunistically combine this data and computing power to model, manage, control, adapt, and optimize virtually any realizable subsystem of interest. Examples of such applications exist in diverse areas ranging from managing extreme events to optimizing everyday processes and improving quality of life. However, the growing data volumes, expanding distribution and dynamism of the data, increasing data heterogeneity and uncertainty about its quality and availability, as well as the growing costs (in time and energy) associated with transporting and processing this data, requires new paradigms and practices in data management and analytics, as well as supporting software stacks before this potential can be realized.

1521-9615/14/$31.00 © 2014 IEEE Copublished by the IEEE CS and the AIP

This special issue of *CiSE* explores the fundamental challenges—as well as the state of the art in solutions—of extreme data, with an attempt on our part to select articles that reflect a balance of methods, experiences, and applications. From innovative algorithmic formulations to implementation frameworks and software stacks, what can accelerate insights from extreme data? End-to-end application workflows and relevant experiences with real applications are of particular interest.

In This Issue

We received a wide variety of interesting articles for this special issue, representing state-of-the-art techniques for managing extreme-scale data in a number of interesting computational science domains, including bioinformatics, astrophysics, and earth sciences (geospatial) applications.

In the first article, "Big Data Applications Using Workflows for Data Parallel Computing," Jianwu Wang and his colleagues make the case that in the era of Big Data, workflow systems must embrace data parallel computing techniques for efficient data analysis and analytics, building on techniques that date back to the 1960s (see, for example, Gul Agha's book *Actors: A Model of Concurrent Computation in Distributed Systems*, MIT Press, 1986). Their approach to build and execute Big Data applications makes use of actor-oriented modeling in data-parallel computing. Two bioinformatics use cases for next-generation sequencing data analysis are presented to demonstrate the effectiveness of the approach being advocated.

In the next two articles, "Ten Years of SkyServer I: Tracking Web and SQL e-Science Usage" and "Ten Years of SkyServer II: How Astronomers and the Public have Embraced e-Science," M. Jordan Raddick and his colleagues working on the Sloan Digital Sky Server astrophysics project present the results obtained by analyzing 10 years' worth of weblog and SQL log data to understand how scientists and the public are using the SDSS's e-Science resources. In part 1, the authors focus on the structure of their extreme-scale database and the Web service architecture to support efficient Web-scale public access in support of the overall argument for how e-Science data providers in general should save all access logs for future analysis. Part 2 focuses on detailed data analysis to demonstrate the impressive reach of SDSS in the broader research community and public at large to show that the methods of e-Science are actually seeing widespread use in other research institutions and education. Could it be that these methods would help to assess the sustainability of scientific software/databases in a longitudinal sense, which was the subject of a recent Supercomputing 2013 workshop (see http://wssspe.researchcomputing.org.uk.)?

Lizhe Wang and his colleagues note in "IK-SVD: Dictionary Learning for Spatial Big Data via Incremental Atom Update" that Big Data is difficult to deal with using traditional methods. Here, the authors examine how to represent a big dataset, which is a fundamental problem in the research of Big Data. Finding an appropriate representation is critical and often requires us to look closely at the domain of interest. This article tackles that challenge by looking at sparse methods to support sampling, reconstruction, compression, retrieval, communication, and classification in remote sensing/imaging applications.

Finally, Matthew Malensek and his colleagues also explore data structuring challenges, focusing on geospatial data in "Evaluating Geospatial Geometry and Proximity Queries Using Distributed Hash Tables." It addresses the challenges associated with supporting geospatial retrievals constrained by arbitrary geometric bounds, geographic proximity, and relevance rankings. The proposed solution involves the use of a lightweight, *distributed* spatial indexing structure—the geoavailability grid. The advantage of a distributed approach is clear here, as it can cope with ever-expanding datasets both in terms of size and the ability to process queries faster. Although distributed systems are known to introduce latency (and a host of other challenges), the authors demonstrate that their approach performs competitively with other spatial indexing technologies.

It's apparent that extreme data will continue to provide new opportunities for insights in all areas of science and engineering—and as such, it will also provide new challenges and opportunities for research. The topic of extreme data promises to remain of significant importance to *CiSE*, and we expect to cover this topic regularly in the coming years. ■

Manish Parashar is a professor in the Electrical and Computer Engineering Department at Rutgers University, he's the director of the Rutgers Discovery Informatics Institute (RDI2) and the US National Science Foundation (NSF) Cloud and Autonomic Computing Center (CAC) at Rutgers, and he's the associate director of the Rutgers Center for Information Assurance. His research

interests focus on applied parallel and distributed computing and computational and data-intensive science and engineering. Parashar has a PhD in computer engineering from Syracuse University. He received the IBM Faculty award twice, as well as a US NSF Career award, and he is an American Assocation for the Advancement of Science (AAAS) and IEEE Fellow. Contact him at parashar@rutgers.edu.

George K. Thiruvathukal is a full professor in the Computer Science Department at Loyola University Chicago, where he also serves as co-director for the Center for Textual Studies and Digital Humanities. His research interests span multiple areas of computer science and interactions with science and the humanities. Thiruvathukal has a PhD in computer science from the Illinois Institute of Technology. He is the editor in chief of *CiSE* and an associate editor for *Computing Now*. Contact him at gkt@cisemagazine.org.

Selected articles and columns from IEEE Computer Society publications are also available for free at http://ComputingNow.computer.org.

ADVERTISER INFORMATION

Advertising Personnel

Marian Anderson: Sr. Advertising Coordinator
Email: manderson@computer.org
Phone: +1 714 816 2139 | Fax: +1 714 821 4010

Sandy Brown: Sr. Business Development Mgr.
Email sbrown@computer.org
Phone: +1 714 816 2144 | Fax: +1 714 821 4010

Advertising Sales Representatives (display)

Central, Northwest, Far East:
Eric Kincaid
Email: e.kincaid@computer.org
Phone: +1 214 673 3742
Fax: +1 888 886 8599

Northeast, Midwest, Europe, Middle East:
Ann & David Schissler
Email: a.schissler@computer.org, d.schissler@computer.org
Phone: +1 508 394 4026
Fax: +1 508 394 1707

Southwest, California:
Mike Hughes
Email: mikehughes@computer.org
Phone: +1 805 529 6790

Southeast:
Heather Buonadies
Email: h.buonadies@computer.org
Phone: +1 973 304 4123
Fax: +1 973 585 7071

Advertising Sales Representatives (Classified Line)

Heather Buonadies
Email: h.buonadies@computer.org
Phone: +1 973 304 4123
Fax: +1 973 585 7071

Advertising Sales Representatives (Jobs Board)

Heather Buonadies
Email: h.buonadies@computer.org
Phone: +1 973 304 4123
Fax: +1 973 585 7071