Milestone Week in Evolving History of Data-Intensive Scalable Computing

By Randal E. Bryant, Carnegie Mellon University and Thomas T. Kwan, Yahoo! Research

The last week of March 2008 saw the emergence of a significant new era in the world of data-intensive scalable computing. Co-sponsored by the Computing Community Consortium (CCC) and Yahoo!, the first ever Hadoop Summit took place on March 25 in Santa Clara, followed by the first Data-Intensive Computing Symposium on March 26 at Yahoo!'s Sunnyvale headquarters.

The Hadoop Summit and Data-Intensive Computing Symposium were the kickoff events of the Big-Data Computing Study Group. Sponsored by the CCC, the study group was formed to foster collaborations between industry, academia, and the U.S. government to advance the state of art in the development and application of large scale computing systems for making intelligent use of the massive amounts of data being generated in science, commerce, and society.

Hadoop Summit

The Hadoop Summit brought together leaders from the Hadoop developer and user community for the first time. (Apache Hadoop, an open source distributed computing project of the Apache Software Foundation, is a distributed file system and parallel execution environment that enables its users to process massive amounts of data.) Originally planned for an audience of 100, the venue was changed to accommodate the enthusiastic response from the open source community. Close to 350 people attended the summit to listen to the talks.

At the summit, Doug Cutting from Yahoo! presented the history of Hadoop and how he started the project, and Eric Baldeschwieler from Yahoo! gave an overview of the Hadoop effort at Yahoo!. (To date, Yahoo! has been the primary contributor to Hadoop.) Various speakers discussed the framework they built atop Hadoop – Kevin Beyer from IBM described the JAQL language, and Chris Olston from Yahoo! described the Pig parallel programming language. Michael Isard from Microsoft Research described DryadLINQ, Microsoft's own language and programming model that bears many similarities to Hadoop. In addition, Andy Konwinski from U.C. Berkeley described using their X-trace tool for monitoring Hadoop performance, and Ben Reed from Yahoo! described the Zookeeper directory and configuration services for Hadoop.

In data management, Michael Stark from Powerset discussed Hbase, a distributed database built atop Hadoop, and Bryan Duxbury from Rapleaf described the application of Hbase for storing pages crawled from the Web. Developers from Facebook described the use of HIVE, a data warehouse built atop Hadoop, and its use at Facebook.

Speakers also presented case studies on the application of Hadoop in various contexts, demonstrating the growing industry acceptance of using Hadoop to solve large-scale, data-intensive problems on highly scalable computing clusters. Case studies were given by speakers from Amazon, Autodesk, Intel/Carnegie Mellon, Yahoo!, and the University

of Maryland, respectively, on using Hadoop to support Amazon Web Services, online search for engineering content, building ground models of Southern California, analyzing web pages, and natural language processing. It became apparent that Hadoop, backed by the power of the open source community, is likely poised to be the default implementation of a parallel computing platform that is rapidly gaining in popularity.

Data-Intensive Computing Symposium

The day after the Hadoop Summit, about 100 researchers from academia, industry, and government laboratories and agencies attended the Data-Intensive Computing Symposium at Yahoo!'s Sunnyvale headquarters. Hosted by Yahoo! and the CCC, the symposium brought together experts in system design, programming, parallel algorithms, data management, scientific applications, and information-based applications to better understand existing capabilities in the development and application of large-scale computing systems, and to explore future opportunities. We co-chaired the symposium, and Bryant opened with a talk contrasting the difference between conventional supercomputers and data-intensive scalable computing (DISC), highlighting the research issues that need to be addressed in the DISC environment.

Experts from several application areas spoke at the DISC Symposium. Alex Szalay from Johns Hopkins discussed the data explosion in astronomy, and how his group has been building data management systems to deal with data issues. Jill Mesirov from the Broad Institute at MIT and Harvard talked about the data explosion in genomic medicine, and the difficulty of replicating scientific experiments that involve the preservation and manipulation of multiple datasets from disjoint data sources. ChengXiang Zhai from the University of Illinois at Urbana-Champaign proposed that Web search application should move towards maximizing personalization, understanding the semantics, and helping users more effectively navigate the information space. Marc Najork from Microsoft Research discussed the mining of large-scale Web graphs, and argued for the need of a theory of the semantics of hyperlinks. Finally, Jon Kleinberg covered the algorithmic perspectives of modeling social processes within large datasets, and how this might influence the design of systems to support online communities. He also raised the issues surrounding the privacy implications of these datasets.

In data management, Joe Hellerstein from U.C. Berkeley discussed the use of declarative specification and dataflow execution of network protocols and distributed systems, and Raghu Ramakrishnan from Yahoo! Research illustrated how his group is prototyping a system that relaxes constraints in traditional database systems to handle scalability and consistency issues in distributed database systems for Web-scale applications.

Speakers from the systems area included Dan Reed from Microsoft Research, Jeff Dean from Google, Garth Gibson from Carnegie Mellon, and Phil Gibbons from Intel Research. Reed emphasized the need to have user experience in mind when designing systems, and challenged systems designers to build simple and easy-to-use tools. Dean described the distributed systems infrastructure at Google, and Gibson provided insights into the unavoidable failure of components in systems at scale and the need to hide the complexity of scale from developers. Gibbons described techniques for improving multicore cache performance and argued for pushing the processing and querying of data to where the sensors are at.

At the symposium, Jeannette Wing from NSF also discussed the agency's broad longer-term interest in data-intensive computing (see her article in the March issue), and Christophe Bisciglia of Google gave an update on the NSF Cluster Exploratory (CluE) program, a partnership among NSF, Google, and IBM. Finally, Ed Lazowska gave a talk over dinner, describing CCC's origin, goals, and activities, and outlined research challenges for the computing field.

Concluding Thoughts

The interest in exploring data-intensive computing by industry is clearly gaining momentum. In addition to the Google/IBM partnership, late last year, Yahoo! announced its partnership with Carnegie Mellon University – the first university to benefit from Yahoo!'s 4,000-processor cluster and expertise in Hadoop. More recently, just a day ahead of the Hadoop Summit, Yahoo! announced an agreement with Computational Research Laboratories (CRL), a subsidiary of Tata Sons Limited in India, where CRL will make the world's fourth fastest supercomputer available to researchers in India for cloud computing research. It is a hopeful sign that others in industry will follow.

Overall, the Hadoop Summit and the DISC Symposium were very well received. At the symposium, Bryant also described that one of the goals of the Big Data Computing Study Group is to recruit around 20 individuals from academia, industry, and government laboratories to serve as advocates for data-intensive computing research. We believe the Hadoop Summit and DISC Symposium were successful in galvanizing a community of practitioners and researchers to help move the field forward, and we look forward to participation from the broader community.