

# Establishing a Big-Data Computing Study Group

Randal E. Bryant, Carnegie Mellon University

Thomas T. Kwan, Yahoo! Research

An ability to quickly and conveniently perform computations over terabyte- and petabyte-scale data sets would enable significant breakthroughs in science, commerce, and other societally important applications. The Big-Data Computing Study Group will explore and enable opportunities for research and applications of high-performance, data-intensive computing systems, benefiting application areas ranging from astronomy to machine translation.

Advances in sensor, networking, computing, and storage technologies have made it possible to collect and store very large (beyond one terabyte) data sets capturing scientific, medical, business, and worldwide web information. Some examples include

- The latest release of the Sloan Digital Sky Survey comprises ten terabytes capturing 287 million unique celestial objects.
- Next-generation DNA microarrays generate around 0.5 terabytes of data to characterize a single genome. Future, personalized medicine will collect this data for individual patients.
- Wal-Mart captures complete point-of-sale data (around 276 million items per day) from each of its 6,000 stores, storing them in a planned four petabyte data warehouse.
- Google's success in the 2005 NIST language translation competition stemmed largely from its use of massive amounts of training data, consisting of 200 billion words of bilingual text and one trillion words of text in the target language English.

How can the capability for computing over large data sets be provided in a way that is cost effective, reliable, and generally usable?

Toward this end, we draw inspiration from the computing systems that have been developed at search engine companies. Their systems collect and organize web documents (a data set of well over one petabyte), and they enable efficient extraction of information from them in response to search queries. These systems are organized as thousands of processing nodes, each consisting of one or more processor cores and one or more disk drives, connected by high-speed, local-area networks. Reliability mechanisms are implemented as part of the system software. Programs can be written in terms of high-level, data-parallel operations, such as the Map/Reduce paradigm introduced by Google, the Pig parallel programming language developed by Yahoo!, and the Dryad framework developed by Microsoft.

There is growing interest in using systems organized along these principles for other data-intensive applications. The open source Hadoop project provides capabilities similar to the Google file system and Map/Reduce. Data centers set up by Yahoo! and by Google and IBM provide access to powerful machines for university researchers, educators, and students. Amazon Web Services makes it possible to rent, rather than buy, large-scale computing and storage. Many challenges remain, however, in fully realizing the potential of this new approach to data-centric computing, ranging from the near-term need of providing Hadoop with multi-user, file system security, to fundamental research problems in system design, parallel algorithms, machine learning, programming languages, and software engineering.

The Big Data Computing Study Group will act to foster research and development in large-scale, data-intensive computing, and to make this capability widely available. It will consist of experts from many areas of computing: storage systems, computer architecture, high-performance computing, databases, programming languages, parallel algorithms, and machine learning. It will also engage experts in application areas that can benefit from new and better ways to create, maintain, and analyze large amounts of data. It will organize workshops, conduct directed studies, and facilitate the development and delivery of educational materials.