Path Forward for Big Data and Extreme Computing

Chaitan Baru, Michael Norman
San Diego Supercomputer Center
UC San Diego, La Jolla, CA

Benchmarks serve multiple purposes—they can assist in debugging systems, testing core functionality, and tuning system performance. Given the broad set of system challenges presented by "big data" applications, we see a need and opportunity for defining big data benchmarks that represent the richness and variety across these applications. Since early 2012, the San Diego Supercomputer Center, through its Center for Large Scale Data Systems Research (CLDS), has fostered a community activity in this area via the *Workshops for Big Data Benchmarking* [WBDB]. These workshops have identified a set of characteristics of big data applications that apply to industry as well as scientific application scenarios. Big data applications are "end-to-end"— typically involving *pipelines* of processing with steps that include aggregation, cleaning, and annotation of large volumes of data; filtering, integration, fusion, subsetting, and compaction of data; and, subsequent analysis, including visualization, data mining, predictive analytics and, eventually, decision making. An outcome of the WBDB workshops has been the formation of a Transaction Processing Council subcommittee on Big Data, which is initially defining a Hadoop systems benchmark, TPC-HS, based on Terasort [TPCBD]. TPC-HS would be a simple, functional benchmark that would assist in determining basic resiliency and scalability features of large-scale systems. Other proposals are also actively under development. One of interest to Big Data and Extreme-scale Computing (BDEC) is the *Deep Analytics Pipeline (DAP),* which defines a sequence of end-to-end processing steps consisting of some of the operations mentioned above. The pipeline model also fits many science applications. Pipeline benchmarks reveal the need for different processing modalities and system characteristics for different steps in the pipeline. For example, early processing steps may process very large volumes of data and may benefit from a Hadoop and MapReduce-style of computing, while later steps may operate on more structured data and may require, say, SMP-style architectures or very large memory systems.

The specification of such data benchmarks will assist the BDEC process, for example, to validate components of the overall system and/or make objective comparison among alternatives designs. It will be important to identify core components of such benchmarks. In addition to the WBDB workshops, others are also interested in this problem. For example, NIST has recently initiated measurement and standards-related activities in Big Data and Data Science [NISTBD, NISTDS], and a group at NERSC has attempted to identify the concepts for data analytics including, basic statistics, graph theory, linear algebra, and others.

Experiences with the Gordon Supercomputer: Deployed in early 2012 at SDSC, Gordon is an NSF-funded supercomputer designed for data-intensive workloads. Key features include 300TB of flash storage local to 64 I/O nodes, 30 TB of node-local flash SSDs, and vSMP software that can aggregate a single addressable memory space across multiple systems. Jobs requiring fast access to large files utilize the Lustre parallel file system. Jobs requiring random IO and high IOPS, e.g. on collections of small files or using databases, benefit from staging the data into flash memory. Recent experience with running whole genome sequencing pipelines demonstrated the advantages of Gordon's architecture. The task required processing 50TB of compressed data using two different 9-step and 5-step pipelines. The task parallelism, memory usage, and amount of data generated varied greatly among different steps of these pipelines. Over the duration of the computation, which required a total of about 36 core-years of computing, the amount of scratch disk space required varied from zero to 257TB and the number of cores used in a single task varied from a few to 5,000 (about 30% of the entire Gordon system).

Comet: The Next Iteration: SDSC was recently awarded a grant from NSF for acquiring and hosting NSF's next supercomputer system, called Comet. Novel aspects of Comet include HPC

virtualization; support for flexible modes of resource allocation, including support for the so-called *science gateways*; and *durable* versus *performance* disk storage. Comet also features substantially more node-local flash SSD storage for data caching than Gordon. *Virtualization* is needed to provide flexibility to enable users to deploy and run their own custom software stacks. Big data applications are characterized by greater diversity in software packages and software stacks, as well as a rapidly evolving software ecosystem—users should be able to run the software that best suits their application. Easy-to-use portals, or *gateways*, that hide the details of the underlying infrastructure from the end user are essential to the success of big data and extreme computing. Given the ubiquitous nature of big data, new communities and new users are coming online rapidly with much higher expectations on user-friendly access to such systems. Finally, there is need for multiple storage systems with different properties tuned to different uses. *Durable storage* in Comet is a backing store to the *performance store.* Other models are possible. For example, a Hadoop-style system may be better suited to the early stages of a pipeline that processes large volumes of data, while later stages may require in-memory filesystem or a data warehouse capability.

New Platforms: The trends in industry, as well as the architecture of systems like Gordon and Comet, point towards a path forward where BDEC systems incorporate a range of capabilities matched up to the differing needs of different parts of a processing pipeline. BDEC systems will be heterogeneous in nature. Currently, big data systems in industry employ shared-nothing architecture with homogeneous, commodity components, to assist in manageability and scalability of the overall system. There is a separation between the active, stream-processing components of the system versus the analytics and *post facto* processing components. However, there is a desire to build "on-line everything" systems, where all processing could be done "inline", *with* the data stream rather than as a post processing step (of course, this still does not obviate the need for analysis of historical data). There is interest in new large memory systems, RAM files, and non-volatile memory systems. New systems will need to support resilient pipeline processing with checkpoint/restart concepts at least between pipeline steps. HPC features such as Infiniband, flash memory, etc. are making their way into commodity systems.

Path forward. A strong lesson learned from big data systems in industry is that the components must be kept fairly simple and robust in order to build highly scalable systems. Thus, a BDEC system should begin with a fairly "straightforward" design and implementation, and iterate further to more complex components and designs using co-design approaches, rather than beginning with a complex system model up front.

A shared community infrastructure for prototyping BDEC system designs would be the first step in making progress. Given multiple agencies interested in such an initiative, e.g. DOE, DARPA, NSF, and others, we recommend a distributed, shared infrastructure where different sites, nodes, components of the distributed infrastructure could be funded by different agencies. The whole would then become greater than the sum of the parts.

URLS:

[NISTBD] NIST Big Data Public Working Group, http://bigdatawg.nist.gov.

[NISTDS] NIST Data Science Symposium, http://www.nist.gov/itl/iad/data-science-symposium-2014.cfm

[TPCBD] TPC Big Data Working Group, http://www.tpc.org/tpcbd/.

[WBDB] Workshops on Big Data Benchmarking, http://clds.sdsc.edu/bdbc/workshops.