4th Workshop on Big Data Benchmarking
http://clds.sdsc.edu/wbdb2013.us

October 09-10, 2013
Brocade Executive Briefing Center
San Jose, CA, USA

Workshop Description

The objective of this workshop is to explore issues in developing industry-standard benchmarks for providing objective measures of the effectiveness of hardware and software systems for big data applications. While micro-benchmarks and functional benchmarks, such as Terasort, play an important role, the workshop is especially interested in application-level, end-to-end benchmarks. A successful benchmark would be simple to implement and execute; cost effective, so that the benefits of executing the benchmark justify its expense; timely, with benchmark versions keeping pace with rapid changes in the marketplace; and verifiable so that results of the benchmark can be validated via independent means. Previous, independent workshops on this topic have led to the development of two benchmark proposals:

- First, based on a *Deep Analytics Pipeline* for event processing, and
- Second, called *BigBench*, based on extending the TPC-DS benchmark with semi-structured and unstructured data and with modified/new queries targeted at those data.

Topics

Related research topics explore a broad range of characteristics that define big data and big data applications, including:

- DATA FEATURES: New feature sets of data including, high-dimensional data, sparse data, event-based data, and enormous data sizes.
- SYSTEM CHARACTERISTICS: Large-scale and evolving system configurations, shifting loads, and heterogeneous technologies for big data and cloud platforms.
- IMPLEMENTATION OPTIONS: Different implementation options such as SQL, NoSQL, Hadoop software ecosystem, and different implementations of HDFS.
- WORKLOAD: Representative big data business problems and corresponding benchmark implementations. Specifying benchmark applications that represent different modalities of big data, including graphs, streams, scientific data, and document collections.
- HARDWARE OPTIONS: Evaluating new options in hardware including different types of HDD, SSD, and main memory, and large-memory systems, and new platform options that include dedicated commodity clusters and cloud platforms.
- DATA GENERATION: Models and procedures for generating large-scale synthetic data with requisite properties.
- BENCHMARK EXECUTIONS RULES, e.g. data scale factors, benchmark versioning to account for rapidly evolving workloads and system configurations, benchmark metrics.
- METRICS FOR EFFICIENCY: Measuring the efficiency of the solution, e.g., based on costs of acquisition, ownership, energy and/or other factors.

- EVALUATION FRAMEWORKS: Tool chains, suites and frameworks for evaluating big data systems.
- EARLY IMPLEMENTATIONS of the Deep Analytics Pipeline or BigBench are solicited. Enhancements to the benchmarks, e.g., incorporation of new data genres and/or new algorithms for machine learning, and alternative benchmark proposals will also be considered.

Paper Submission, Review, and Publication
Papers should be formatted using the Springer LNCS Proceedings format.
Selected papers will be published in Lecture Notes in Computer Science by Springer-Verlag.

Full papers: max. 12 pages
Short papers: max. 6 pages

Please use the following submission system to submit your paper:
https://cmt.research.microsoft.com/WBDB2013/

Submission Dates:
- September 6, 2013: Due date for full workshop papers submission
- September 20, 2013: Notification of paper acceptance to authors
- October 09-10, 2013: Workshop
- October 18, 2013: Camera-ready of accepted papers

Previous workshops:
- 1st WBDB: May 2012, San Jose, CA, USA (http://clds.ucsd.edu/wbdb2012)
- 2nd WBDB: December 2012, Pune, India (http://clds.ucsd.edu/wbdb2012.in)
- 3rd WBDB: July 2013, Xi'an, China (http://clds.ucsd.edu/wbdb2013.cn)

Workshop Chair(s):
- Chaitan Baru (San Diego Supercomputer Center, UC San Diego)
- Milind Bhandarkar (Pivotal)
- Tilmann Rabl (Middleware Services Research Group, U of Toronto)

Publicity Chair:
- Florian Stegmaier (University of Passau, Germany)

Program Committee
- Dhruba Borthakur (Facebook)
- Yanpei Chen (Cloudera)
- John Galloway
- Ahmad Ghazal (Oracle)
- Boris Glavic (IIT Chicago)
- Bhaskar Gowda (Intel)
- Eyal Gutkind (Mellanox)
- Songlin Hu (Chinese Academy of Science)
- Jian Li (IBM)
- D. K. Panda (Ohio State University)
- Scott Pearson (Brocade)
- Meikel Poess (Oracle)
- Francois Raab (InfoSizing)
- Jerry Zhao (Google)
- Jianfeng Zhan (Chinese Academy of Sciences)

Contact: For questions please contact Chaitan Baru (baru[at]sdsc[dot]edu)