Big Data Benchmarking

Chaitan Baru
San Diego Supercomputer Center
UC San Diego, USA
baru@sdsc.edu

Milind Bhandarkar
Greenplum
EMC, USA
milind.bhandarkar@emc.com

Raghunath Nambiar
Cisco Systems, Inc.
USA
rnambiar@cisco.com

Meikel Poess
Oracle Corporation
USA
meikel.poess@oracle.com

Tilmann Rabl
Middleware Systems Research Group
University of Toronto, Canada
tilmann.rabl@utoronto.ca

ABSTRACT

We provide a summary of the outcomes from the *Workshop on Big Data Benchmarking (WBDB2012)* held on May 8-9, 2012 in San Jose, CA. The workshop discussed a number of issues related to big data benchmarking definitions and benchmark processes, and was attended by 60 invitees representing 45 different organizations from industry and academia. Attendees were selected based on their experience and expertise in one or more areas of big data, database systems, performance benchmarking, and big data applications. The participants concluded that there exists both a need and an opportunity for defining benchmarks to capture the end-to-end aspects of big data applications. The metrics for such benchmarks would need to include metrics for performance as well as price/performance, and consider several costs including total system cost, setup cost, and energy costs. The next *Workshop on Big Data Benchmarking* is scheduled to be held on December 17-18, 2012 in Pune, India.

1. INTRODUCTION

The world has been in the midst of an extraordinary information explosion over the past decade, spurred by the rapid growth in the use of the Internet and the number of connected devices worldwide. We are experiencing a rate of increase in data growth that is faster than at any point throughout history. Enterprise application data as well as machine-generated data continue to grow exponentially, challenging industry experts and researchers to develop new innovative techniques to evaluate and benchmark hardware and software technologies and products. Studies have estimated that the total amount of enterprise data will grow from about 0.5 zettabyte in 2008 to 35 zettabytes in 2020 [1].

This phenomenon is global in nature, for example, with Asia rapidly emerging as a major source of data users as well as data generators. With increasing penetration of data-driven computing, Web and mobile technologies, and enterprise computing, the emerging markets have the potential for further adding to the already rapid growth in data. The development of benchmarks is a necessary step in helping quantify system architectures designed to confront this data deluge and tackle big data applications. The First Workshop on Big Data Benchmarking (WBDB2012) [2] was a first important step towards the development of such benchmarks, for providing objective measures of the effectiveness of hardware and software systems dealing with big data applications. The benchmarks would facilitate evaluation of alternative solutions and provide for comparisons among different solution approaches by characterizing the new feature sets, enormous data sizes, large-scale and evolving system configurations, shifting loads, and heterogeneous technologies of big-data and cloud platforms.

The objective of WBDB2012 was to identify key issues and to launch an activity around the definition of reference benchmarks that can capture the essence of big data application scenarios. The invited attendees included academic and industry researchers and practitioners with backgrounds in big data, database systems, benchmarking and system performance, cloud storage and computing, and related areas. Each invitee was required to submit a two-page abstract on their vision towards Big Data Benchmarking. The program committee reviewed the abstracts and classified them in to four categories namely: data generation, benchmark properties, benchmarking process and hardware and software aspects for big data workloads.

2. WORKSHOP DESCRIPTION

The workshop was held on May 8-9, 2012 in San Jose, CA, hosted by Brocade at their Executive Briefing Center. There were a total of about 60 attendees representing 45 different organizations. The workshop structure was similar on both days of the meeting, with presentations in the morning sessions followed by discussions in the afternoon. Each day started with three 15-minute talks. On the first day, these talks provided background on industry benchmarking efforts and standards and discussed desirable attributes and properties of competitive industry benchmarks. The 15-minute talks on the second day focused on big data applications and the different genres of big data, such as genomic and geospatial data. On both days, these opening presentations were followed by 20 "lightning talks" of 5 minutes

MBDS'12, September 21, 2012, San Jose, CA, USA.
Copyright © 2012 ACM 978-1-4503-1752-8... $15.00.

each by the invited attendees. The lightning talk format proved to be extremely effective for putting forward and sharing key technical ideas among a group of experts. The workshop group of about 60 attendees was then arbitrarily divided into two equal groups for the afternoon discussion session. Both groups were provided the same topic outline for discussion and ideas from both groups were collated at the end of the day.

The workshop program committee was responsible for drawing up the list of invitees; issuing invitations; designing the workshop structure described above; grouping white papers into lightning talk sessions; and developing the outline to guide the afternoon discussions.

3. BACKGROUND

There was general consensus at the workshop that a big data benchmarking activity should begin at the end-application level, by attempting to characterize the end-to-end needs and requirements of big data applications. While isolating individual steps of such an application such as, say, sorting, is also of interest, this should still be done in the context of the broader application scenarios. Furthermore, a range of data genres need to be considered for big data including, for example, structured, semi-structured, and unstructured data; graphs; streams; genomic data; array-based data; and geospatial data. It is important to model the core set of operations relevant to each genre while also looking for similarities across genres. Interestingly, it may be possible to identify relevant application scenarios that involve a variety of data genres and require a range of big data processing capabilities. An example discussed at the workshop was that of data management at an Internet-scale business, say, for an enterprise like Facebook or Netflix. One could construct a plausible use case for such applications that requires big data capabilities for managing data streams (click streams), weblogs, text sorting and indexing, graph construction and traversals, as well as some geospatial data processing and structured data processing.

It is also possible that a single application scenario may not plausibly capture all the key aspects, for genres as well as operations on data that may be broadly relevant to big data applications. Thus, there may be a need for multiple benchmark definitions, based on differing scenarios, which together would capture the full range of possibilities.

There are good examples of successful benchmarks by consortia such as the Transaction Processing Performance Council (TPC) and Standard Performance Evaluation Corporation (SPEC); from industry leaders, such as VMMark and Top500; and benchmarks like Terasort and Graph500 for specific operations and data genres (social graphs), respectively. It is fair to ask whether a new big data benchmark could simply be built upon existing benchmark efforts, by extending the current benchmark definitions appropriately. While this may be possible, a number of issues need to be considered including, whether:

- the existing benchmark models relevant application scenarios;
- the existing benchmark can be naturally and easily scaled to the large data volumes that will be required for big data benchmarking;
- such benchmarks can be used more or less “as is”, without requiring significant re-engineering to produce data and queries (operations) with the right set of characteristics for big data applications; and
- the benchmark has inherent restrictions or limitations, such as, say, a requirement to implement all queries to the system in SQL.

Several extant benchmarking efforts were mentioned and discussed such as, the Statistical Workload Injector for MapReduce SWIM, developed at the University of California, Berkeley [3]; GridMix3, developed at Yahoo! [4]; YCSB++, developed at the Carnegie Mellon University on top of YCSB of Yahoo! [5]; and TPC-DS , the latest benchmark addition to the TPC’s suite of decision support benchmarks [6].

4. CONCLUSION

The Workshop on Big Data Benchmarking held on May 8-9, 2012 in San Jose, CA discussed a number of issues related to big data benchmarking definitions and processes. The workshop concluded that there was both a need and an opportunity to define benchmarks for big data applications. Such benchmarks would model end-to-end application scenarios and consider a variety of costs, including total system cost, setup cost, and energy costs. Benchmark metrics would include performance metrics as well as cost metrics. Several next steps are underway. An informal “big data benchmarking community” has been formed. Biweekly phone conferences are being used to keep this group engaged and to share information among members. A few members of this community are beginning to work on simple examples of end-to-end benchmarks. Efforts are underway to obtain funding to support pilot benchmarking activity.

The next Big Data Benchmarking workshop will be held on December 17-18, 2012 in Pune, India, hosted by Persistent Systems. A third workshop is being planned for July 2013 in Xi’an, China.

5. REFERENCES

[1] Anual Digital Universe Report, 2010, EMC Corp. http://www.emc.com/digital_universe.

[2] Workshop on Big Data Benchmarking. 2012. http://clds.ucsd.edu/wbdb2012

[3] Statistical Workload Injector for MapReduce (SWIM): https://github.com/SWIMProjectUCB/SWIM/wiki

[4] Gridmix3: git://git.apache.org/hadoop-mapreduce.git/src/contrib/gridmix/

[5] Patil S. et al.: YCSB++ : Benchmarking and Performance Debugging Advanced Features in Scalable Table Stores, SOCC ’11

[6] Raghunath Othayoth Nambiar, Meikel Poess: The Making of TPC-DS. VLDB 2006: 1049-105