Towards an Industry Standard for Benchmarking Big Data Workloads

Raghunath Nambiar

Cisco Systems, Inc.
3800 Zanker Road
San Jose, CA 95134, USA

rnambiar@cisco.com

Industry Standard Benchmarks

Historically, robust and meaningful benchmark standards have been crucial to the advancement of the computing industry. Without them, assessing relative performance between disparate vendor architectures is virtually impossible. Demands for audited and verifiable benchmarks have existed since buyers were first confronted with a choice between purchasing one piece of hardware over another, and have been driven by their desire to compare price and performance on an apples-to-apples basis. Over the years, benchmarks have proven useful to both systems/software vendors and purchasers. Vendors use benchmarks to demonstrate performance competitiveness for their existing products and to improve/monitor performance of products-under-development; in addition, many buyers reference benchmark results when considering new technologies and products. Finally, benchmarks help vendors improve their products through competition [1].

The two most prominent industry standard benchmark organizations are the Transaction Processing Performance Council (TPC) and Systems Performance Evaluation Corporation (SPEC). The TPC's primary focus is total system performance under a database workload, including: hardware, operating system, and I/O system. All results have a price-performance metric audited by an independent TPC certified auditor. Like the TPC, SPEC develops suites of benchmarks to in measure system performance, packaged with source code and tools and are extensively tested for portability before release. Unlike the TPC, SPEC results are peer audited.

Big Data Landscape

The information explosion and technological revolution have significantly changed the world that live in a great extent. Today, thirty percent of world population has internet access. There are fifteen billion devices connected to the internet that is more than two devices for every human being living on the planet earth. If Facebook, were a country, with 900 Million unique users, it would be the third largest in the world between India and the United states. The McKinsey Global Institute estimates that data volume is growing 40% per year, and will grow 44x between 2009 and 2020 [2] estimated to 35 Zettabytes.

Web 2.0 companies have done great handling big data problem by embracing open source framework like Apache Hadoop and using massive scale-out configuration of low-end servers to store, manage and process massive sets of structured, semi structured and unstructured data.

1. The industry and technology landscape is changing quickly and enterprise companies are also challenged with big data problems. This is fueled by the increased adaption of internet and connected devices in everyday business. Businesses are realizing that they can benefit from big data and analytics. Historically,

they made business decisions based on transactional data stored in relational database management systems like CRM and ERP applications. Now, mining nontraditional data sources like Call Detail Records, machine logs, weblogs, social media, customer feedback, sensor data, email, pictures and video along with traditional enterprise data can improve their business process, improve user experience and gain competitive edge.

Call for a New Standard

Though there are several innovations, growing collection of new technologies, and product in the big data space, enterprises are challenged with the influx of new data sources and these new technologies and new products, and also "several claims". This calls for industry standards for evaluation, comparison and characterization of technologies, products and workloads in terms of performance, cost of ownership and energy efficiency. Manageability is another important consideration.

Industry standard bodies like Transaction Processing Performance Council (TPC) and Systems Performance Evaluation Corporation (SPEC) have been well serving the industry with various standards, but to address the big data challenge, new standards need to be evolved, and existing standards need to be improved to complement the new standards.

The situation is not a lot different from when the TPC was originally founded in 1998 in response to a growing trend in "benchmarketing," or attempt by vendors to publish questionable benchmark results in order to increase sales. The need for a vendor-neutral standards organization that focused on creating and administering fair and comprehensive benchmark specifications to objectively evaluate database systems under demanding, but consistent and comparable workloads, quickly became apparent. Several influential database academics and industry leaders began working to establish an organization charged with leading the effort to impose order and consistency to the process of benchmarking products fairly and objectively – this effort ultimately culminated in the formation of the TPC.

The author believes that the Workshop on Big Data Benchmarking (WBDB 2012) is a first important step towards the development of a benchmark standard that measures the effectiveness of hardware and software systems dealing with big data applications, and will be used by vendors, researcher and end-users. Vendor point of view these standards define a level playing field to competitive comparisons and enable monitoring release to release progress of hardware and software products. Like other industry standard benchmarks, these standards are expected to help the end user customers to compare performance, cost of ownership and energy efficiencies of big data platforms. Standardize workloads and metrics like this will also help researchers to develop and enhance relevant technologies for big data applications.

A Framework for the New Standard

Developing an industry standard benchmark is a complex task that includes the development of business model and workload representative of common big data applications, execution rules that ensure important aspects like repeatability, and metrics that is simple and understandable by both technical and business people. Key components of an industry standard benchmark and considerations are listed in Table 1.

Table 1: Key Components of industry standard benchmark

Components	Data Generation	Workload and Execution Rules	Metric
Considerations	Generation of large dataset that is relevant (structured, semi structured, unstructured) Consistent across multiple hardware and software platforms Can be generated in a timely manner	Representative of several common Big Data use cases Technology and product agnostic Exercise relevant hardware (compute, storage, network) and software Repeatable (minimal run to run variations) Practical (Workload can be represented on practical hardware configurations)	Relevant in terms of performance, Capex and Opex Simple and understandable by vendors, researchers and end users Verifiable (independent audit or peer audit)

The author envisions that the development of this benchmark workload will follow an open source model, in which users are allowed to experiment, enhance and contribute to the benchmark workload development. Once the workload is mature, it will be available for existing industry standard committees like the TPC, to adapt and develop specifications that meets their requirements (including execution rules, metrics, and methodologies for pricing and energy efficiency) while the development in the open source community continues.

References

[1] R. Nambiar, M. Lanken, N.Wakou, F. Carman, M. Majdalany: Transaction Processing Performance Council (TPC): Twenty Years Later – A Look Back, A Look Ahead: R. Nambiar, M. Poess (Eds.): Performance Evaluation, Measurement and Characterization of Complex Systems: LNCS vol. 6417, Springer 2011, ISBN 978-3-642-18205-1

[2] Big data: The next frontier for innovation, competition, and productivity: McKinsey Global Institute 2011