Industry Standard for Benchmarking Big Data Systems

NIST Big Data Public Working Group
IEEE Big Data Workshop
October 27, 2014

Raghunath Nambiar
Cisco Distinguished Engineer
Chief Technologist Big Data and Analytics
Chairman TPC Big Data Committee
rnambiar@cisco.com
Big Data is Transforming the World

- Big Data has become an integral part of enterprise IT ecosystem across major verticals including agriculture, education, energy, finance, healthcare, insurance and media services
- Big Data is one of the most talked about topics in government and research
  - Challenges continue to be the 5V’s
  - Becoming center of 3I’s – Investments, Innovation, Improvization

The Big Data technology and services market represents a fast-growing multibillion-dollar worldwide opportunity (Source: IDC 2013)

Big Data technology and services market will grow at a 27% compound annual growth rate (CAGR) to $32.4 billion through 2017 - or about six times the growth rate of the overall information and communication technology (ICT) market (Source: IDC 2013)

Big Data will drive $240 billion of worldwide IT spending in 2016 directly or indirectly (Source: IDC 2013)

74% of organizations have invested in or plan to Invest in Big Data. 30% have already invested in big data technology, 19% plan to invest within the next year, and an additional 15% plan to invest within two years using (Source: Gartner 2014)

United States federal agencies are estimated to spent $6 billion in 2014 and then to $8 billion by 2017 at a compound annual growth rate of 10 percent (Source: Biometrics Research Group 2013)
Top Challenge for Enterprise Customers

What platform to pick in terms of performance, price-performance, and energy efficiency?

Source: IDC 2014
Not Easily Verifiable Claims and Chaos

There are Claims (not discrediting them) but not easily variable or comparable due to lack of standards.
Remember the 1980’s?

State of the Nature - early 1980’s

The industry began a race that has accelerated over time: automation of daily end-user business transactions. The first application that received widespread focus was automated teller transactions (ATM), but we’ve seen this automation trend ripple through almost every area of business, from grocery stores to gas stations. As opposed to the batch-computing model that dominated the industry in the 1960’s and 1970’s, this new online model of computing had relatively unsophisticated clerks and consumers directly conducting simple update transactions against an on-line database system. Thus, the on-line transaction processing industry was born, an industry that now represents billions of dollars in annual sales.

Early Attempts at Civilized Competition

In the April 1, 1985 issue of Datamation, Jim Gray in collaboration with 24 others from academia and industry, published (anonymously) an article titled, “A Measure of Transaction Processing Power.” This article outlined a test for on-line transaction processing which was given the title of “DebitCredit.” Unlike the TP1 benchmark, Gray’s DebitCredit benchmark specified a true system-level benchmark where the network and user interaction components of the workload were included. In addition, it outlined several other key features of the benchmarking process that were later incorporated into the TPC process.

The TPC Lays Down the Law

While Gray’s DebitCredit ideas were widely praised by industry opinion makers, the DebitCredit benchmark had the same success in curbing bad benchmarking as the prohibition did in stopping excessive drinking. In fact, according to industry analysts like Omri Serlin, the situation only got worse. Without a standards body to supervise the testing and publishing, vendors began to publish extraordinary marketing claims on both TP1 and DebitCredit. They often deleted key requirements in DebitCredit to improve their performance results.

From 1985 through 1988, vendors used TP1 and DebitCredit—or their own interpretation of these benchmarks—to muddy the already murky performance waters. Omri Serlin had had enough. He spearheaded a campaign to see if this mess could be straightened out. On August 10, 1988, Serlin had successfully convinced eight companies to form the Transaction Processing Performance Council (TPC).
TPC-Big Data Standard Initiative

- Big Data was identified as one of the top areas for industry standard benchmark developments at the VLDB 2014, TPCTC 2014, WBDB 2014 and other conferences
- Continuing TPC’s commitment to developing relevant benchmark standards
- TPC-BD Working Group formed in October 2013 to evaluate big data workload(s) and make recommendations to the TPC general council
- TPC-BD Subcommittee formed in February 2014 to develop an Express benchmark based on already popular TeraSort workload
- In July 2014 TPCx-HS became industry’s first standard for benchmarking Big Data Systems
- TPC to continue work on other benchmark(s)
TPCx-HS Benchmark

• x: Express, H: Hadoop, S:Sort
• Provides verifiable performance, price/performance, general availability, and optional energy consumption metrics of big data systems
• Enable measurement of both hardware and software including Hadoop Runtime, Hadoop Filesystem API compatible systems and MapReduce layers
• Primary audience is enterprise customers (not public clouds)
Workload

- Based on TeraSort workload
  - TeraSort is part of Apache Hadoop distribution.
    org.apache.hadoop.examples.terasort
- A valid run consists of five separate phases run sequentially
- The benchmark test consists of two runs and run with lower metric is reported
- No configuration or tuning changes or reboot are allowed between the two runs
Scale Factors

- The TPCx-HS follows a stepped Scale factor model (like in TPC-H and TPC-DS)
- The test dataset must be chosen from the set of fixed Scale Factors defined as follows:
  - 1TB, 3TB, 10TB, 30TB, 100TB, 300TB, 1000TB, 3000TB, 10000TB.
  The corresponding number of records are
  - 10B, 30B, 100B, 300B, 1000B, 3000B, 10000B, 30000B, 100000B,
  where each record is 100 bytes generated by HSGen
- The TPC will continuously evaluate adding larger Scale Factors and retiring smaller Scale Factors based on industry trends
TPCx-H Energy Metric

• The TPCx-HS energy results are expected to be accurate representations of system performance and energy consumption, based on TPC-Energy Specification

• TPCx-HS Energy metric reports the power per performance and is expressed as Watts/HSph@SF:

\[
\frac{E}{(T \times HSph@SF)}
\]

where E is the energy consumption for the reported run, T is the elapsed time in seconds for the reported run, and HSph@SF is the performance metric
Positive Feedback from Media

Over two dozen articles

- Register: http://www.theregister.co.uk/2014/08/20/big_data_benchmark_from_tpc/
Developing an industry standard benchmark for a new environment like Big Data has taken the dedicated efforts of experts across many companies. Thanks to:

Andrew Bond (Red Hat), Andrew Masland (NEC), Avik Dey (Intel), Brian Caufield (IBM), Chaitanya Baru (SDSC), Da Qi Ren (Huawei), Dileep Kumar (Cloudera), Jamie Reding (Microsoft), John Fowler (Oracle), John Poelman (IBM), Karthik Kulkarni (Cisco), Meikel Poess (Oracle), Mike Brey (Oracle), Mike Crocker (SAP), Paul Cao (HP), Reza Taheri (VMware), Simon Harris (IBM), Tariq Magdon-Ismail (VMware), Wayne Smith (Intel), Yanpei Chen (Cloudera), Michael Majdalany (L&M), Forrest Carman (Owen Media) and Andreas Hotea (Hotea Solutions).
Outlook ...

We envision that TPCx-HS will be a useful benchmark standard to buyers, as they evaluate new systems for Hadoop deployments in terms of performance, price/performance and energy efficiency. And for vendors in demonstrating competitiveness of their products ...