Conference proceedings these days are full of performance comparisons of various data management technologies using "*variations*" of industry standard benchmarks. The VLDV 2010 proceedings[1] contain 25 papers that used TPC benchmarks in their performance characterizations (the count of 25 does not include papers submitted to the TPC Technical Conference track of VLDB). The theme for nearly all these papers was Big Data, proving the demand for industry standard benchmarks in this community. However, the current usage of TPC benchmarks in this area fails in multiple regards:

- It goes without saying that none of the TPC benchmarks was designed specifically for the 2012 vintage Big Data. We need new benchmarks.
- Almost all the papers cited a variation of a TPC benchmark: Only some queries were used; the database was not scaled according to performance; code that was supposed to be run as is was modified; etc. The whole value of industry-standard benchmarks is how one can use them to compare results produced by different parties. When a sponsor takes liberties with the workloads or the execution rules, this value evaporates.
- It is well-known that TPC guards its benchmarks against unauthorized rule. These benchmarks are TPC's main IP, its brand name. Then how could 25 papers cite unaudited, unofficial results? This was all done under the academic use guidelines that allow using unaudited results for academic purposes. But it was interesting to observe the questions that the attendees posed to presenters, or evaluation of one set of results by authors of a rival set. Accusations of improper use, cutting corners, unfair comparisons, etc. reminded the author of the Wild West days of 1980s when the many benchmark wars resulted in the creation of the SPEC and TPC consortiums. The academic use guidelines were put in place for occasional use in non-controversial studies. Clearly, we need standard benchmarks with detailed execution rules and some form of review for benchmarks to find widespread use, even for academic purposes.

But what model should we follow for industry-standard benchmarks for big data? Let us look at a few existing models:

- TPC benchmarks have been *the* database performance standards for nearly 25 years. The reasons are:
    - + TPC benchmarks have longevity. TPC-C has been in existence for 20 years, and new results still get published. TPC-C is the TPC benchmark most requested by users even after 20 years.
    - + Easy to compare: due to audit requirements and strict, detailed run rules, one can compare results published by two different entities
    - + Scaling: this is a property that should be of utmost importance to the Big Data community. TPC benchmarks are just as meaningful at the high-end of the market as at the low-end; as relevant on clusters as on single servers. Furthermore, one cannot get around the demands of the workload by throwing more of some resource (memory, CPU, etc.) at the problem statement to make it go away.

    On the other hand, there exist major problems with the TPC benchmarks:

    - - They are very hard and expensive to run. An official, audited result is impossible to sponsor for all but a few very large corporations. This model is unworkable in the Big Data market.

[1] (Proceedings of VLDB 2010, 36th International Conference on Very Large Data Bases, Singapore: 13 to 17 Sept 2010, 2010)

    - TPC does not provide benchmarking kits. One has to develop a kit, or acquire a kit provided by a major H/W or S/W vendor, tethering one to that vendor.
    - Most software used in TPC disclosure have *DeWitt Clauses*, banning disclosure of performance data without the consent of the software vendor. This is not inherent in the TPC benchmark specifications; but is a fact of life since almost all results are with only a few DBMS packages (due to the two issues above).

- SPEC got its start as the developer of hardware benchmarks, but over the years has seen success in many more areas. The reasons are:
    + SPEC provides the benchmarking kit. Anyone can buy the kit (costing a few 100 to a few 1,000 US Dollars) and run the benchmark.
    + SPEC benchmarks make good use of open source software, making it possible for performance analysts to test a system without having to partner with other companies, and without the binds of DeWitt clauses.
    + SPEC updates each benchmark every few years. A side effect, important to the Big Data workshop, is that updates mean that one does not have to worry about a benchmark that has to be relevant for 20+ years. If it is good enough for 3-5 years, that's long enough until replacement by the next rev.
    + SPEC produces a large number of benchmarks, and develops them quickly.
    + SPEC benchmarks can run in *base* and *peak* modes. If one publishes peak results, one is required to also supply the base results.

  The issues with SPEC benchmarks are:
    - Database benchmarking has not been SPEC's forte.
    - SPEC benchmarks have traditionally been run as is, using their kits. Some of the newer SPEC benchmarks allow the test sponsor to insert components of the sponsor's choosing in the software stack. How should one compare results measured on commercial software to results on open source components?

- An often-forgotten set of industry-standard benchmarks are those put out by vendors to test their own products. VMmark from VMware and SAP SD are two good examples of this. The strength of these benchmarks are:
    + They are developed by a single entity. The development process is much faster, and less controversial.
    + The run rules are simpler, but are augmented as needed by the company if they discover incidents of sponsors gaming the benchmark.
    + The owner of the benchmark is in effect also the auditor of the results. Even though the results are not audited by an (expensive) impartial auditor, they are blessed by the final authority in the application and the benchmark.
    + Each benchmark is intended for a specific purpose, and works quite well to that end. There is no reason to worry about questions such as "What if someone comes up with a new technology that breaks the benchmark? Therefore, this benchmark has to be designed a priori to deal with all new technologies." Vendor-supplied benchmarks measure only one thing, but measure it well.

The issues are obvious:

- VMmark and SAP SD measure only specific software stacks. No SD results for Oracle E-Business Suit, or VMmark results for Hyper-V.
- Can we really trust a single commercial entity with an industry-standard benchmark?

So looking at the best and worst properties of the benchmarks above, I propose the following set of desirable properties for a Big Data benchmark:

- Provide the kit. A specification-only benchmark means its official usage will be limited to deep-pocket entities. It will also mean a specification that has to predict the many ways that the benchmark can be *gamed*, and have detailed wording to disallow it. A kit (a complete benchmarking kit, not just app software) makes that easy.
- The reference kit will run against open source software. If a test sponsor wants to insert a different, commercially-available component in the stack, the sponsor will be required to also disclose the results with the base, open source reference components.
- Scaling is paramount. A 1K-node cluster and a 10-node cluster should not be solving the problem.
- Don't develop a single benchmark to satisfy many markets; even if that means disappointing people who want to compare many different technologies using a single benchmark. The wider the net we cast, the more impossible it will be to develop a benchmark that is relevant, or that can be developed in a reasonable amount of time. It is better to have two or three simpler, non-comparable benchmarks.
- Limit academic use to new projects that use the benchmark as an internal tool. Any comparison to other results, even by researchers, requires a reviewed benchmark result. (I recognize this will be unpopular; but it's an issue that needs to be discussed.)
- Develop with an end-of-life in mind.

About me: I have been doing system performance analysis for 30 years, first at Bell Labs, then for 20 years at HP, and the last 5 at VMware. I have focused on tuning large systems, typically running databases and transaction processing workloads. I first published a TPC benchmark result in 1990, too many to count since. I chair the TPC-V virtualization benchmark development subcommittee.