Big Dada Benchmarking Workshop

San Jose, CA May 8-9, 2012

Paul Denzinger, Hewlett-Packard

Proposal for a Big Data Benchmark Repository

Abstract:

The benchmarking space for Big Data covers a wide range of use-case scenarios having very different performance characteristics that cannot be easily addressed by a few targeted benchmarks, as has been done formerly for OLTP and Data Warehousing systems.

Big Data can benefit from adopting a community-based benchmark repository, similar in function to open-source product development. Such a repository would become an active collection point for benchmark frameworks and results, with contributors free to submit benchmarks of their choice, thus extending the range and scope of benchmarks available to the community.

We at HP have recently started to develop our Big Data Strategy, and my particular interest in this area is one of performance characterizations of Big Bata systems. One of our main goals at HP is to ensure that our systems are optimized and balanced in such ways as to deliver the intended performance to meet the processing requirements of our customers.

Customer usage patterns of our systems spans the entire range and scale of big data related applications, as well as the numerous emerging data base products. Our customers have expectations that we as their vendor will engineer systems that perform well and efficiently with workloads from basic Hadoop map/reduce jobs or with any number of NoSQl/NewSQL products running advanced analytics, regardless of the use-case scenarios. It is from this perspective that we are initiating efforts to begin our performance characterizations of our systems and solutions.

Big Data challenges cover a very wide range of application domains, data types and modalities, and use-case scenarios. I would argue that the problem of benchmarking in this space is considerably more challenging than with OLTP systems and the more traditional Data Warehouse systems where usage patterns were more narrowly constrained. Because of this, I am suspect that a few benchmarks in any particular area of Big Data will be near sufficient for characterizing the many different types of workloads that exist presently, much less those which will certainly be developed in the future.

This is not to diminish any of the efforts being done to benchmark these systems – in fact, these are certainly needed. Whereas some parties may be interested in pursuing very specific types of

benchmarks, we at HP are interested more the entire range of benchmarks available for use. We very much would like to have a repository of sorts cataloguing Big Data benchmarks as they are developed and made available to the Big Data community. I can envision something similar to the TPC organization's list of benchmarks, but more extensive, and populated by contributors from the Big Data community – without a controlling committee – covering any domain of Big Data processing. Consider it as open source for Big Data benchmarks.

This Big Data benchmark repository should include at least the following elements:

- A name, description and a discussion of the use-cases for which the benchmark is intended
- Access to the code comprising the benchmark, pertinent configuration information, workload drivers, and any applicable monitoring tools/connectors
- Guidelines to setup, configurations and usage, test objectives, etc.
- Benchmark results, including hardware/software platforms & versions, details on client sessions, jobs, concurrency, throughputs, utilizations, latencies – all of the typical performance metrics

A simple example of this is the YCSB benchmark framework. It includes many of the components listed above, and includes details for customizing the benchmark to specific back-end DBs. Users are free to modify the tests to their own requirements as well as extend the tests to cover additional test cases.

Such a repository would serve as a valuable tool for all interested parties in the Big Data community, whether from vendors, the private sector, academia, or elsewhere. Creators of new benchmarks would have a common location for easy dissemination. The benchmarks themselves would evolve and be refined and extended much like open source products today. Interested parties could use these benchmarks to test products or ideas, and subject to their discretion, make the results available to community via the same repository.

We at HP could better serve our customers by having access to (as well as contributing) what would become a large and diverse set of use-case scenarios that would more closely match the performance characteristics exhibited by our customer's systems, enabling us to produce more effective solutions. Furthermore, any company would have free access to any results published in the repository, giving them valuable insight into the performance expectations from various workloads and configurations.