Revisiting Issues in Benchmark Metric Selection

Christopher Elford, Dippy Aggarwal(✉), and Shreyas Shekhar

Intel Corporation, Santa Clara, USA
{chris.l.elford,dippy.aggarwal,shreyas.shekhar}@intel.com

Abstract. In 1986, Fleming and Wallace presented a case advocating the use of geomean in benchmark results. 23 years later in 2009, Alain Crolotte followed up on that proposal at TPCTC using TPC-D as a reference. Now 11 years later it is time to present another perspective on the age-old argument regarding the best metric for summarizing benchmark data. The aim of this paper is two-fold: (1) summarize the definition and interpretation of the current benchmark metrics for the OLAP family of the TPC benchmarks, including TPC-H, TPC-DS, and TPCx-BB. (2) illustrate the impact and tradeoffs of different statistical measures on the overall benchmark metric score, using both conceptual and data-driven arguments. Our hope is that the paper reinvigorates interest in the benchmark community to re-evaluate the design of benchmark metrics and offer insights that can influence the future direction of benchmark metrics design.

Keywords: Benchmarks · Databases · Performance · Metrics

1 Introduction

Database benchmarking is a valuable process that measures combined performance of various components of a database engine along with the underlying hardware. The investment of time and money is motivated by the desire to identify areas for optimizations and to demonstrate one's standing against competition in the market. As laid out by Peter Mattson, General Chair of MLPerf, "Benchmarking aligns research with development, engineering with marketing, and competitors across the industry in pursuit of a clear objective [1]."

TPC [21] (Transaction Processing Performance Council), a consortium led by a committee of industry and academic experts, is responsible for developing, managing, and auditing several database benchmarks spanning the areas of analytics, big data, transaction processing, IoT, and AI. This paper focuses specifically on the three analytical/Big Data benchmarks provided by TPC: TPC-H [2], TPC-DS [3], and TPCx-BB [4].

One of the key challenges in developing a database benchmark is the definition of a composite benchmark metric which serves as an objective criteria/score for database vendors and system designers to evaluate and compare

R. Nambiar and M. Poess (Eds.): TPCTC 2020, LNCS 12752, pp. 35–47, 2021.
https://doi.org/10.1007/978-3-030-84924-5_3

their products against competitors. The metric captures end to end database performance by combining individual performance and interactions of different scenarios/phases such as data load, power run (simulating a single user test with all queries in the benchmark run in a sequential order), throughput run (multiple query streams running concurrently), and others. The performance of these individual phases is combined in specific ratios using different statistical measures such as arithmetic or geometric mean and the choice of a particular statistic influences database vendor's decision for where to focus their efforts to drive performance optimizations and hence have the highest impact on the benchmark metric score.

While we cover the implications of employing arithmetic vs. geometric mean in the context of benchmark metric later in the paper (Sect. 2), we discuss them briefly here. Using arithmetic mean to summarize the performance (runtime) of the power run of a benchmark implies that optimizing only the few top running queries is enough to boost the benchmark score while ignoring the short running queries since the higher execution times will skew the average towards them. However, geometric mean by its definition offers equal weightage to all queries.

There have been discussions in the past which highlight the design reasons that have led to the current metric definitions of some of the performance benchmarks [5,6,8,10,12]. In 1986, Fleming and Wallace [6] proved that geometric mean is the only correct average of normalized measurements. Fast forward 20 years, Crolotte [5] proved that arithmetic mean is the only valid metric to summarize single-stream elapsed times for decision-support benchmark. Crolotte cited pitfalls from using geomean in TPC-D to motivate the omission of geomean in TPC-DS which is based on TPC-D. Nambiar et al. [8] justify their choice of using arithmetic mean for the power run component of TPC-DS benchmark by highlighting that more real world customer scenarios are around optimizing long running queries. Citron et al. [12] conducted an interesting analysis around the topic of appropriate mean for comparing multi-workload groups of metrics to each other based on a literature survey and their independent experiments. The authors observe that the "best" choice of statistical average fluctuates widely from changing even a single ingredient workload in the group. They conclude that relative computer performance cannot be assessed in an absolute manner by a small number of contrasting or toy benchmarks regardless of the choice of statistical average. John [10] presents a case for weighted average. As another independent observation derived from reviewing of the current TPC-DS publications, longest query accounts for 7.6% of the power run time but 12.5% of the Throughput runtime. This presents another argument in favor of focusing the optimization efforts on the top long running queries since that will have an even larger impact on the throughput metric as well. Literature survey on the topic of benchmark metrics is replete with arguments favoring the use of one mean vs. other [9,11].

McChesney [13] and Mashey [9,11] provides a primer on the relationship between lognormal data and geometric means. Recently, in a journal (Communications in Statistics - Theory and Practice), Vogel summarized the deep history of a number of data distributions and ways to summarize central tendency

across multiple disciplines. It also supports the premise that there is a relationship between lognormal data and the geometric mean.

Our motivation to revive this discussion comes from following three observations: (1) TPC-DS consists of 99 queries and by using arithmetic mean to summarize the results from complete run of these 99 queries, we believe that we are not only steering the attention of database vendors to a handul of top running queries but it also begs the question for why do we need the rest of queries in the benchmark if performance optimizations around them are not effectively counted towards benchmark score. (2) Literature has numerous examples of using geomean for lognormal data and interestingly, TPC-DS does reflect lognormal distribution [11] (details in Sect. 3). (3) While TPC-DS consists of 99 queries representative of industry operations, not all 99 queries would necessarily be relevant to all end users. If a long running query is irrelevant to a given end user the metric is of limited utility.

Through this paper, we hope to revive the discussion and present our observations highlighting limitations in the current definition of the TPC-DS metric which is based on using arithmetic means to capture results of the power run phase of the benchmark. We raise the following questions, answers to which will lay the foundation for our ideas in this paper.

1. What constitutes the objective criteria/benchmark metric? How do the current metrics of TPC ad-hoc query benchmarks compare?
2. Analyze the impact and tradeoffs of different statistical measures on the benchmark metric score. Is the Benchmark Metric "Reward Structure" for the newer benchmarks better than the earlier benchmarks?
3. How hard is it to design a metric that "fairly" rewards optimization? What influences the constituent mix of a benchmark metric?

We address the first two questions in detail in Sect. 2 and the challenges and pros and cons of different metric designs is covered in Sect. 3. Based on the recognized issues, we then leverage the discussion to present our proposal of considering the use of geomean for computing the performance of the power run component of TPC-DS. Section 4 summarizes our ideas along with enumeration of questions that must be addressed for the successful implementation of our proposal of employing geomean for TPC-DS power run phase.

Before moving forward with the discussion, below are the definitions of the various terms that will reoccur throughout the paper and represent key ingredients used in defining the metric.

1. *Benchmark "Reward Structure"*: How much does the benchmark score improve on optimizing a system ingredient?
2. *Power Run*: A single series of queries run one at a time (in a defined order)
3. *Throughput Run*: Several concurrent series of queries run together (each series in a defined order)
4. *Load*: Creation of the database (includes population, indexing, etc.)
5. *Maintenance*: Updating [refreshing] database with a [partial] fresh data snapshot
6. *Scale Factor*: Any of a series of discrete allowed database sizes

2 Survey of TPC OLAP/Big Data Metrics

In this section, we analyze the characteristics of three analytical benchmarks offered by TPC: TPC-H, TPC-DS, and TPCx-BB, focusing on both their commonalities and differences (e.g., the presence of a data maintainence phase in TPC-H and TPC-DS which is not a part of TPCx-BB) and comparing their respective benchmark metrics. All the three benchmarks involve running power and throughput run phases. We draw interpretations of these benchmark metrics to highlight the design rationale behind each of them and how the they steer system's and database vendors attention to optimize certain phases of their products.

2.1 TPC-H

TPC-H has been the most popular and widely embraced analytical benchmark across both academic research (7500+ results for TPC-H in google scholar [14]) and industry practitioners (290+ publications [15]). It is a decision-support benchmark which analyzes large volumes of data to answer ad-hoc queries representing business critical scenarios such as pricing summary report which captures the amount of business that was billed, shipped, or returned. A TPC-H performance metric (Queries per hour - Qph) for a given Scale factor (SF) consists of two geometrically equal weighted phases:

1. Power@Size is the combined duration of 22 sequential queries and two refresh queries. The queries are given geometrically equal weights. Q(i,0) in the second equation represents read only queries while R(j,0) correspond to the two refresh queries.

$$QphH@Size = \sqrt{Power@Size * Throughput@Size} \tag{1}$$

$$TPC-HPower@Size = \frac{3600 * SF}{\sqrt[24]{\prod_{i=1}^{i=22} Q(i,0) * \prod_{j=1}^{j=2} R(j,0)}} \tag{2}$$

2. Throughput@Size is the combined duration of multiple simultaneous streams of 22 queries (and their associated refreshes).

Interpretation

The metric definition above can be interpreted as follows.

1. There is no need to optimize database load time in TPC-H.
2. There is no bias toward optimizing data maintenance (update queries) any more or less than the read only analytics queries.

In the next section, we do a similar analysis of the metric design of another TPC benchmark, TPC-DS which is gaining popularity over the past few years since it supersedes TPC-H in terms of schema complexity and the query coverage (99 queries in TPC-DS vs. 22 in TPC-H) [8].

2.2 TPC-DS

TPC-DS exhibits several commonalities with TPC-H given that both are decision support, analytical benchmarks. They have several phases in common including power run, throughput run, and data maintainence but there also exists key areas where it diverges from the TPC-H benchmark design. The throughput run in TPC-DS is run twice, once immediately after the power run and once after data maintainence. The TPC-DS performance metric significantly differs from TPC-H in that for a given scale factor (SF) consists of four geometricly equal weighted phases:

1. Power (T_{PT}) is the combined duration of 99 sequential queries. The value is multipled by the stream count to place it into a similar scale as the Throughput timing.
2. Throughput (T_{TT}) is the combined duration of multiple simultaneous streams of 99 queries running twice.
3. Maintenance (T_{DM}) is the time to perform data maintenance (update queries) which is run in two phases.
4. Load (T_{LD}) is the time to load the database, configure indices, etc. It constitutes one percent of the duration multipled by the stream count, S_{CT}.

$$QphDS@SF = \frac{SF * S_{CT} * 99}{\sqrt[4]{T_{PT} * T_{TT} * T_{DM} * T_{LD}}} \tag{3}$$

Interpretation

1. Because a geomean is used to combine the four phases, ingredient providers are encouraged to equally optimize for all four activities (single-stream, multi-stream, data updates, load efficiency). For example, given a choice between improving single stream performance by 10%, multiple stream performance by 10%, data maintenance by 10%, or data load by 10%, all would provide an equal bump to the benchmark metric.
2. Within the power metric, total time is used which rewards the provider from optimizing the longest running queries and minimizes benefits from optimizing shorter running queries. This begs the question of why TPC-DS includes all 99 queries.
3. While the power run metric and throughput components of TPC-H include both read-only and refresh queries, TPC-DS incorporates refresh queries in a separate phase (T_{DM}) and accounts for only 99 read-only queries for the power and throughput run components of its metrics (T_{PT}, T_{TT}).

The use of arithmetic mean vs. geometric mean in the TPC-H and TPC-DS metrics respectively for the power run brings back the historical controversary: *is it really desirable to give an equal benefit to the benchmark score from improving a short running query and a long running query*? For example, the use of geometric mean in TPC-H (Eq. 2) offers equal weightage to a 10% optimization observed

over a long running query vs. a short running one whereas the arithmetic mean used in TPC-DS favors optimizations to long running queries.

In the next section, we cover yet another TPC benchmark, TPCx-BB which is an analytical benchmark like TPC-H and TPC-DS but extends the workload beyond the traditional, structured data supported by TPC-H and TPC-DS.

2.3 TPCx-BB

TPCx-BB extends the previous analytical benchmarks by integrating machine learning and natural language processing queries as well alongside SQL queries on structured data. The scale of supported database sizes in TPCx-BB is also higher (1 PB) compared to H/DS (100 TB). A TPCx-BB performance metric consists of two geometrically equal weighted phases coupled with a portion of the load time.

1. T_{LD} is 10% of the load time.
2. T_{PT} is computed much like TPC-H with a geomean of the 30 TPCx-BB query durations.
3. T_{TT} is computed much like TPC-H with a net time of multiple streams running simultaneously.

$$BBQpm@SF = \frac{SF * 60 * 30}{T_{LD} + \sqrt[2]{T_{PT} * T_{TT}}} \tag{4}$$

$$T_{PT} = 30 * \sqrt[30]{\prod_{i=1}^{i=30} Q(i,0)} \tag{5}$$

Interpretation

1. Unlike TPC-H and TPC-DS, the current generation of TPCx-BB has no refresh queries to include in the metric.
2. Because only 10% of the load time is accounted for in the metric, the load time needs to be significantly longer than the power/throughput run time before load time optimization is justified.

While this section provided a conceptual and formula based interpretation of the current metrics, in the next section we present a quantitative evidence for the interpretations drawn above for each of the benchmark metrics.

2.4 Survey Summary

We now illustrate how the optimizations over different benchmark phases impact the benchmark metric using actual numbers from published reports on each of the three benchmarks. We took a current TPC-H [16], TPC-DS [17], and TPCx-BB [18] publication, and their associated Load, Query, and Refresh times. We then artificially adjusted by 10% their load times, refresh times, power run and

throughput run query times, and their longest and shortest query times. For simplicity, for the longest and shortest query time adjustment, we reduced the respective query duration in both the power and in the throughput runs. This is likely a conservative estimate for the impact on the Throughput Run because interaction effects will likely inflate times in queries running in parallel in other streams.

We then recomputed a theoretical benchmark score based on the adjusted times. In the table below (Table 1) we show the impact to overall score from a theoretical 10% optimization to each portion of the benchmark described as follows. The first row in the table serves as the baseline.

- *Load 10% faster*: Adapts the load time to 90% of the original time listed in the publication.
- *All RF Queries 10% faster*: Computes the benchmark score by revising the runtime of each of the refresh queries to 90% of their original runtimes in the publication.
- *All Power/Throughput Run Queries 10% faster*: Computes the benchmark score by revising the runtime of all power and throughput run queries to 90% of their original runtimes in the publication.
- *All Power Run Queries 10% faster*: Computes the benchmark score by revising the runtime of all power run queries to 90% of their original runtimes in the publication.
- *All Throughput Run Queries 10% faster*: Computes the benchmark score by revising the runtime of all throughput run queries to 90% of their original runtimes in the publication.

Table 1. TPC-H, TPC-DS, TPCx-BB optimization reward structure

Row#	Workload element	TPC-H	TPC-DS	TPCx-BB
1	Base	100.0%	100.0%	100.0%
2	Load 10% faster	na	102.7%	100.5%
3	All RF queries 10% faster	100.4%	102.7%	n/a
4	All power/throughput run queries 10% faster	110.6%	105.4%	110.5%
5	All power run queries 10% faster	104.9%	102.7%	105.1%
6	All throughput run queries 10% faster	105.4%	102.7%	105.1%
7	Longest query 10% faster	100.6%	100.5%	101.4%
8	Shortest query 10% faster	100.2%	100.0%	100.2%

Observations

1. As shown in Table 1 , the three workloads intentionally or unintentionally reward/place very different importance on data load and data maintenance (Rows 2,3). In contrast to TPC-H, TPC-DS dramatically increases the importance of optimizing the load and refresh phases; giving them equal metric impact as the power and throughput runs (Rows 5,6). In practice, some end

users may heavily stress data load time by frequently refreshing data sets while other users may only infrequently reload their data sets. Similarly some users will heavily rely on data maintenance while others may not.
2. With respect to an across the board optimization of all queries in the power and/or throughput run phases (Rows 4–6), Table 1 shows that all three TPC benchmarks agree and actively try to provide reward to optimizing whole benchmark phases. But the award percentage in TPC-DS (105.4%) is almost half compared to the incentive provided in TPC-H and TPCx-BB (110.6% and 110.5% in TPC-H and TPCx-BB respectively). In practice, some users may stress a single query at a time while other users stress multiple concurrent queries.
3. As expected, TPC-H significantly rewards optimization of the whole power or throughput run phase (Rows 4–6), gives a reasonable reward to optimizing the longest running query (Row 7), and gives a tiny reward to optimizing refresh queries (Row 3) or the shortest running query (Row 8).
4. TPC-DS maintains TPC-H's reasonable reward granted to optimizing the longest running query but gives essentially no benefit to optimizing the shortest running queries (Row 8).
5. TPCx-BB is similar to TPC-H with significant rewards to optimizing the whole power or throughput run phase, a modest reward to optimizing the longest running query, and a tiny reward to optimizing the shortest running query.

In summary, TPC-H and TPCx-BB both recognize that the short running queries still play an important role and use a geomean which prevents them being completely ignored. In contrast, the design of the TPC-DS metric encourages the short running queries to be ignored during optimization. This is ironic given that one of the interesting claims to fame of TPC-DS is that it dramatically increases the count of real-world user query types relative to workloads like its TPC-H predecessor. However, the number of queries that actually make a significant difference to the metric may not be nearly as large in TPC-DS.

For all three of these benchmarks, TPC and its members invested years of engineering effort to identify a variety of real world ad-hoc queries that represent the types of queries performed in decision support and big data production environments. Having closed on a set of queries, each benchmark formed a metric that encourages test sponsors to optimize selected elements of the solution. We posit that all three benchmarks contain limitations or implicit assumptions in their selection and will discuss it in the next section as it relates to each of the thought experiments in Table 1 above.

3 Metric Design Considerations

Fleming and Wallace [6] passionately argue for geomean for summarizing normalized results and provides a proof based on certain assumptions that geomean is decidedly better. Crolotte [5] argues with equal passion for arithmetic mean and provides a similar proof based on alternate assumptions that arithmetic

mean is decisively better. He also provides an interesting example outlining how the choice of metric contributed to the early demise of TPC-D. Our argument for a choice of metric while no less passionate is conceptually based on how TPC benchmarks are designed.

The choice of metric fundamentally guides the one publishing the benchmark towards system ingredients that should be optimized to achieve higher scores. In Sects. 1 and 2, we compared and contrasted the ingredients that are encouraged for optimization in the TPC-H, TPC-DS, and TPCx-BB benchmarks. In this section, we lay the foundation for our proposal to consider geometric mean and its variations in defining the power run component of the TPC-DS metric.

3.1 Ideal Metric and Challenges

From the perspective of an end user downloading a TPC benchmark result and its disclosure and trying to understand how to use the result in their decisions on how to provision their environment, the answer is fairly simple: *The end user would like a metric that rewards the particular queries that the user tends to use in* ***their*** *environment highly and ignores the remaining queries.* If, for example, those happen to all be short running queries then the short running queries should be highly rewarded. The perspective that different end users may stress different queries highlights the fact that there are challenges inherent in defining an ideal benchmark metric that can serve the use cases for different database vendors and system designers without incurring any bias. We recognize the following specific challenges in the design of an ideal benchmark metric.

1. Since each user may have a different set of queries that they are interested in, it is practically impossible to create a single metric that reflects all combinations. All three benchmarks largely ignore the relative frequency of the queries.
2. Since there are dependencies amongst the queries such as earlier queries loading data that will then be used by later queries and more complex interaction effects between streams in the Throughput run, it is impossible in the general case to reliably extract timing for one or two queries from a TPC benchmark run and assume that they would represent the performance of those queries in isolation.
3. A savvy user can create a metric using data extracted from TPC benchmarks that represents their environment but that representation may or may not be an adequate substitute for actually running the query mix that they are interested in situ.

3.2 Candidate Solutions

We offer two proposals for revising the TPC-DS power run metric subcomponent in the overall benchmark score.

Using Geometric Mean Instead of Arithmetic Mean: For this option, we will posit a variant of TPC-DS that equally rewards a given percentage speedup

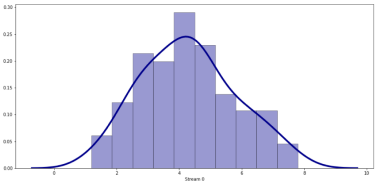
to each query independent of its power duration (similar to TPC-H and TPCx-BB). Note the difference in the revised power run metric component in Eq. 7 (geomean) compared to T_{PT} described in Sect. 2.2.

$$GphDS@SF = \frac{SF * S_{CT} * 99}{\sqrt[4]{G_{PT} * T_{TT} * T_{DM} * T_{LD}}} \tag{6}$$

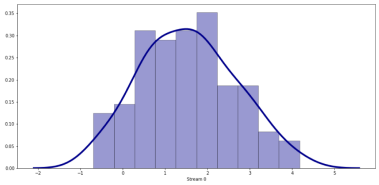
where

$$G_{PT} = S_{CT} * 99 * \sqrt[99]{\prod_{i=1}^{i=99} Q(i,0)} \tag{7}$$

[7,11,13] posit that data of a lognormal distribution tends to be best summarized via a geomean. Despite the concerns of [5], the authors of this paper believe that the geomean has an important role to play to ensure that short and long queries each can play their role within the benchmark and can serve as an alternative to a weighted mean when it is "too hard" to agree on relative query weights. In Figs. 1a and 1b we show two TPC-DS publications [19,20] which are randomly chosen from the available publication results. The bell shaped graphs confirm how the power run query times are indeed lognormal further supporting the position that geomean does have a role to play if we want to reflect queries across the spectrum.



(a) Cisco publication [19]



(b) Alibaba publication [20]

Fig. 1. Demonstrating lognormal characteristic of TPC-DS

Using Weighted Geometric Mean Instead of Arithmetic/Pure Geometric Mean: For this second option, we will provide some targeted weighting to give a larger reward for optimizing certain queries over others. While all three benchmarks contain a number of real world representative ad-hoc query use cases, none of the three benchmarks define the real-world relative frequency of each query. If they did, a weighted average could be used to accurately reflect the importance of the queries. As an example, if a short running query in practice is run 1,000,000 times more often than a long running query it could be realistically accounted for in the metric. Unfortunately agreeing on the frequency of execution of all the query types within the TPC benchmark development committees would be an even harder process than agreeing on the list of queries to include in the first place. We discuss this further in Sect. 4.

$$WGphDS@SF = \frac{SF * S_{CT} * 99}{\sqrt[4]{WG_{PT} * T_{TT} * T_{DM} * T_{LD}}} \tag{8}$$

where

$$WG_{PT} = S_{CT} * 99 * \sqrt[\sum_{j=1}^{n} w_j]{\prod_{i=1}^{i=99} Q(i,0)^{w_i}} \tag{9}$$

Table 2 shows the results of applying these alternative metric schemes (geomean and weighted geomean) to the TPC-DS benchmark score. For discussion, we will maintain the current TPC-DS award structure as is and assume that in the prescence of an across the board optimization to all queries in the power and/or throughput run it should gain a similar benchmark score boost. We refer to the query runtimes from the same TPC-DS publication [17] that was referenced for results shown in Table 1 but instead of using an arithmetic mean, we use alternative metrics in Table 2.

Table 2. Demonstrating experimental optimization reward structure

Workload element	Current metric	Pure Geomean	Weighted Geomean
Base	100.00%	100.00%	100.00%
Load 10% faster	102.67%	102.67%	102.67%
All RF queries 10% faster	102.67%	102.67%	102.67%
All power/throughput run queries 10% faster	105.41%	105.41%	105.41%
All power run queries 10% faster	102.67%	102.67%	102.67%
All throughput run queries 10% faster	102.67%	102.67%	102.67%
Longest query 10% faster	100.47%	100.31%	100.29%
Shortest query 10% faster	100.00%	100.03%	100.04%

In the current metric (Column 2), we recognize that while a 10% optimization on the longest query is able to improve the score to 100.47% but the same amount of optimization over the shortest running query shows no improvement to the overall score. Leveraging geomean (Column 3) instead helps to alleviate that skew by offering some small incentive to optimizations on short running queries as well (100.03%) by slightly reducing the impact of long running queries. but the relative impact of the long running queries is still quite high. The weighted geomean in the Column 4 illustrates an approach to balance this further by offering a higher reward to the short running queries. While actual query weights should be determined by the committee, for the sake of discussion of the impact of a weighted geomean we use an arbitrary scenario that gives 4× weight to the the shortest running 25 queries. Further adjusting the weights could produce a more significant adjustment as desired.

4 Conclusions and Next Steps

This concludes our discussion around the issues, tradeoffs, and potential improvements in the benchmark metrics design. We summarized the definition and interpretation of the current benchmark metrics for the OLAP family of the TPC benchmarks, including TPC-H, TPC-DS, and TPCx-BB. Using that as the foundation, we illustrated the impact and tradeoffs of different statistical measures on the overall benchmark metric score, using both conceptual and data-driven arguments. While the two TPC analytical benchmarks, TPC-H and TPCx-BB employ a geometric mean to summarize the impact of power run on the overall benchmark score, TPC-DS uses an arithmetic mean. Our results demonstrate how a weighted geometric mean can offer additional control for defining the benchmark metric as opposed to arithmetic or geometric means which are either skewed towards long running queries or completely eliminate any ranking among the queries in terms of their respective runtimes.

As the next step, we would like to invite the TPC committee, industry leaders and practitioners in the field of performance engineering and benchmarking to consider re-evaluating the reward structure used in the design of benchmark metrics . Some of the aspects to consider are:

1. Which queries should have the largest reward?
2. Which phases should impact benchmark score the most?
3. Should TPC-DS 3.0 incorporate a weighted geomean? If so, how should the query weights be defined (e.g., query frequency, query duration, resource use such as CPU/IO, etc.).

Having presented these ideas for future discussions, we also recognize the complexity involved in implementing concrete solutions. To address the challenge, one of the steps can be to turn our attention back to the fundamental question that forms the basis of this discussion of the relevant statistical measure for a benchmark metric. The question being, *what are the goals that the benchmark designers seek to achieve and how does the current choice of metrics (and underlying statistical measures) help those goals?* We believe there are two main expectations.

1. The metric offer insights to the database and system designers on where to focus their optimization efforts to improve their benchmark scores.
2. As a benchmark designer, we want to steer the vendors' attention to areas that will improve end user experience and solve real challenges in the industry.

The choice of statistical measures in the existing metrics is based solely on the query runtimes. The arithmetic mean focuses one's efforts primarily on optimizing the "long" running queries while geomean offers equal incentives towards all queries. We believe that in our discussion around this area, we should look beyond query runtimes to include the aspect of representativeness of the query (e.g., based on either its frequency of use or some other resource usage) in the real-world customer uses cases. One opportunity for future work is to create a taxonomy regarding resource types suitable for inclusion in weighting decisions.

We hope that bringing forth and reviving this discussion around benchmark metrics selection serves as a worthwhile effort to not only ensure that TPC benchmarks continue to offer valuable insights to customers and database and system designers alike but also address how we can make end users more inclined to embrace results.

References

1. ML Benchmark Design Challenges - Hot Chips. https://www.hotchips.org/hc31/HC31_1.9_MethodologyAndMLSystem-MLPerf-rev-b.pdf. Accessed 14 Sept 2020
2. TPC-H. http://www.tpc.org/tpch/default5.asp. Accessed 14 Sept 2020
3. TPC-DS. http://www.tpc.org/tpcds/default5.asp. Accessed 14 Sept 2020
4. TPCx-BB. http://www.tpc.org/tpcx-bb/default5.asp. Accessed 14 Sept 2020
5. Crolotte, A.: Issues in benchmark metric selection. In: Nambiar, R., Poess, M. (eds.) TPCTC 2009. LNCS, vol. 5895, pp. 146–152. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-10424-4_11
6. Fleming, P.J., Wallace, J.J.: How not to lie with statistics: the correct way to summarize benchmark results. Commun. ACM **29**(3), 218–221 (1986)
7. Vogel, R.M.: The geometric mean?. Commun. Stat. Theor. Methods 1–13 (2020)
8. Nambiar, R.O., Poess, M.: The Making of TPC-DS. In: VLDB, vol. 6, pp. 1049–1058 (2006)
9. Mashey, J.R.: War of the benchmark means: time for a truce. ACM SIGARCH Comput. Archit. News **32**(4), 1–14 (2004)
10. John, LK.: More on finding a single number to indicate overall performance of a benchmark suite. ACM SIGARCH Comput. Archit. News **32**(1), 3–8 (2004)
11. Iqbal, M.F., John, L.K.: Confusion by all means. In: Proceedings of the 6th International Workshop on Unique Chips and Systems (UCAS-6). (2010)
12. Citron, D., Hurani, A., Gnadrey, A.: The harmonic or geometric mean: does it really matter? ACM SIGARCH Comput. Archit. News **34**(4), 18–25 (2006)
13. Three simple statistics for your data visualizations. https://breakforsense.net/three-statistics/. Accessed 14 Sept 2020
14. TPC-H Google Scholar Search Results. https://scholar.google.com/scholar?as_vis=1&q=tpc-h+&hl=en&as_sdt=1,48. Accessed 14 Sept 2020
15. TPC-H Results. http://www.tpc.org/tpch/results/tpch_advanced_sort5.asp?PRINTVER=false&FLTCOL1=ALL&ADDFILTERROW=&filterRowCount=1&SRTCOL1=h_sponsor&SRTDIR1=ASC&ADDSORTROW=&sortRowCount=1&DISPRES=100+PERCENT&include_withdrawn_results=none&include_historic_results=yes. Accessed 14 Sept 2020
16. TPC-H Publication. http://www.tpc.org/tpch/results/tpch_result_detail5.asp?id=119040201. Accessed 14 Sept 2020
17. TPC-DS Publication. http://www.tpc.org/tpcds/results/tpcds_result_detail5.asp?id=120061701. Accessed 14 Sept 2020
18. TPCx-BB Publication. http://www.tpc.org/tpcx-bb/results/tpcxbb_result_detail5.asp?id=119101101. Accessed 14 Sept 2020
19. Cisco UCS Integrated Infrastructure for Big Data. http://www.tpc.org/tpcds/results/tpcds_result_detail5.asp?id=118030501. Accessed 14 Sept 2020
20. Alibaba Cloud AnalyticDB. http://www.tpc.org/tpcds/results/tpcds_result_detail5.asp?id=120061701. Accessed 14 Sept 2020
21. TPC. http://www.tpc.org. Accessed 15 Sept 2020