Issues in Benchmark Metric Selection

Alain Crolotte

Teradata Corporation, 100 N Sepulveda Blvd.
El Segundo, Ca. 90045
alain.crolotte@teradata.com

Abstract. It is true that a metric can influence a benchmark but will esoteric metrics create more problems than they will solve? We answer this question affirmatively by examining the case of the TPC-D metric which used the much debated geometric mean for the single-stream test. We will show how a simple choice influenced the benchmark and its conduct and, to some extent, DBMS development. After examining other alternatives our conclusion is that the "real" measure for a decision-support benchmark is the arithmetic mean.

1 Introduction

The purpose of this paper is to examine a basic problem facing benchmark designers when selecting a metric in the context of a decision-support benchmark. Once the database has been populated and the queries defined comes the apparently simple task to define a metric i.e. a single number that will summarize the elapsed times. The most natural way to accomplish that is to use the arithmetic mean. But, since no benchmark participant has run all the queries at this stage of the game, the usual nagging question comes up: "What if one query dominates the entire set?" To solve the outlier problem, although it has not appeared yet, a potential solution would be to define a rule such as "throw away one" by which benchmark participants would be allowed to remove their worst query time from the final result set.

Another approach to the problem of aggregating highly skewed observations could be to select a different metric that would hopefully use all the queries without being dominated. In this paper we examine potential ways to define *a priori* a metric to summarize a set of raw observations when potentially large discrepancies could occur in the value set as is potentially the case in a decision-support benchmark. In particular we will look into the choice made by the subcommittee who designed the TPC-D benchmark in tackling with this problem. We will also show that the only valid *a priori* metric for a decision-support benchmark is the arithmetic mean.

The paper is organized in 6 sections including this introduction. In section 2 we provide background information on the problem of choosing a metric. Section 3 examines potential solutions to the general problem while section 4 looks at the actual metric used by the TPC-D and examines some of the consequences of this choice. In section 5 we make the case for the arithmetic mean being the only valid alternative for a single-stream decision-support metric. Finally, a short section summarizing the conclusions of this study is provided.

R. Nambiar and M. Poess (Eds.): TPCTC 2009, LNCS 5895, pp. 146–152, 2009.

2 Background

In the case of TPC-D, the subcommittee in charge of defining the benchmark chose an interesting approach for the definition of the single-stream metric based on the geometric mean but with a twist. The TPC-D benchmark specification including the metric is described in [1]. The TPC-D benchmark (now obsolete and replaced by the TPC-H benchmark) was defined in the early nineties and, at this time, the geometric mean enjoyed a deserved popularity as a metric in benchmarks. This popularity can be traced to a paper in which the authors showed that the only correct metric to summarize normalized benchmark results is the geometric mean [2]. The key here is "normalized". The numbers to be summarized in the case of a single-steam decision-support benchmark such as TPC-D are elapsed times i.e. raw numbers and therefore not normalized. In [5] for instance, it is argued that the arithmetic mean should be used when averaging times. Therefore the use of the geometric mean in this case cannot be justified on this basis unlike for the Spec benchmark (see [3] for a discussion on the relative merits of the arithmetic mean, geometric mean and harmonic mean in this context).

Aside for the so-called lack of sensitivity to outliers the geometric mean has an interesting property established in section 3. It treats all relative changes in the same way – for instance, if an observation varies by 10% the relative change in the geometric mean is the same whether the observation is large or small. The arithmetic mean has the same property but only for "absolute" improvements.

The subject of finding a single measure of performance has been debated for quite a while in the area of computer performance (see [2], [6], [7] and [8]) but the issue is eventually resolved by paying attention at the type of data is under review. The particular subject of decision-support metrics has been examined in [4].

3 Characteristics of Central Tendency

There are three basic metrics that are commonly used, (1) the arithmetic mean or simple average, (2) the geometric mean and (3), the harmonic mean – they are usually associated with basic operations namely addition, multiplication and division In this paper we will place things in a more general context. First we start with a set of n positive numbers (elapsed times) x_1, x_2, up to x_n. We assume that these observations will be weighted equally in all cases. We usually denote the arithmetic mean as m, the geometric mean as g and the harmonic mean as h. The usual formulas for these metrics are

$$m = \frac{1}{n}\sum x_i \tag{1}$$

$$g = \left(\prod x_i\right)^{1/n} \tag{2}$$

$$h = \frac{1}{\frac{1}{n}\sum \frac{1}{x_i}} \tag{3}$$

Given a monotonic function ϕ one can define a measure of central tendency associated with this function as follows (see also [4] and [9]) called phi-average:

$$\phi(M_\phi) = \frac{1}{n}\sum_{i=1}^{n} \phi(x_i) \tag{4}$$

The formulas for the usual metrics can be represented this way using the function $\phi(x)=x^r$ hence we will name the phi-average associated with r-th power function the r-th power average. The arithmetic mean is obtained for r=1, the harmonic mean for r=-1 and the geometric mean is obtained as a limit case for r=0. Taking the logarithm in both sides of equation 2 we obtain an equivalent expression for the geometric mean;

$$\log g = \frac{1}{n}\sum \log x_i \tag{5}$$

To see how this formula is a limit case of equation 4 with $\phi(x)$ being x to the power r when r tends toward zero use the fact that

$$x^r = e^{r \log x} = 1 + r \log x + o(r) \tag{6}$$

Assume that all observations in (5) are constant except say x_k and take the derivative on both sides. We get $dg/g=(1/n)(dx_k/x_k)$ and the results still holds approximately when the differentials dg and dx_k are replaced by finite quantities. This establishes the geometric mean property mentioned in section 2 that a relative variation in an observation will result in the same overall variation of the geometric mean whether the observation is small or large.

Also, it can readily be established that the r-th power average is an increasing function of r so that the geometric mean is always lower than the arithmetic mean [2]. Whether we look at formula 2 or formula 5 we see that there is a very undesirable property of the geometric mean. If just one observation is equal to zero the geometric average of all the quantities is equal to zero. In other words the geometric mean puts overwhelming emphasis on small observations in cases where large or regular values are mixed with very small values. In order to solve the problem and as suggested in [4] we could use any r-th power average with r between 0 and 1 – for instance r=1/2. The corresponding formula for the one-half power average s is

$$s = \left(\sqrt{x_1} + \sqrt{x_2} + \ldots \sqrt{x_n}\right)^2 \tag{7}$$

While s avoids the pitfall of the discontinuity for small values exhibited by the geometric mean and reduces the influence of outliers in the high end, it is not familiar

like the harmonic mean, the geometric mean, the arithmetic mean or even the 2-nd power or square power average which is so popular in Statistics. Also, it does not have meaning in terms of the problem at hand since it has the dimension of a time (like the geometric mean) but it is not a time (also like the geometric mean).

Another way to deal with the basic problem exhibited by the geometric mean with observations close to zero is to avoid those observations by adding a small positive quantity to all observations e.g. 1/1000 or 1/100. We can define a log-based mean within the framework of the ϕ-average by using the function $\phi(x)=\log(x+a)$ where a is a fixed positive integer. This new measure of central tendency g_a is called the a-displaced geometric mean and is given by the formula:

$$\log(g_a + a) = \frac{1}{n}\sum \log(x_i + a) \tag{8}$$

If a is very small, this new average has "almost" all the interesting properties of the geometric mean and avoids its unpleasant pitfall. In [4] we have shown that, for a given set of observations, the a-displaced geometric mean is an increasing function of a. So for any positive a it will be always larger than the geometric mean g. Also the arithmetic mean is a limit case for the a-displaced average when a becomes very large. Factoring a out of all quantities under the log in equation 8 and remembering that, when x is small log(1+x) is equivalent to x leads to (9) when a tends toward infinity:

$$\log a + \log(1 + \frac{g_a}{a}) = \log a + \frac{1}{n}\sum \log(1 + \frac{x_i}{a}) = \log a + \frac{1}{a}\frac{1}{n}\sum x_i + o(\frac{1}{a}) \tag{9}$$

Therefore ga/a tends towards 0 and

$$\log(1 + \frac{g_a}{a}) = \frac{g_a}{a} + o(\frac{1}{a}) = \frac{m}{a} + o(\frac{1}{a}) \tag{10}$$

Therefore, when a becomes large, the a-displaced average tends toward the arithmetic mean. In summary the a-displaced mean is an increasing function of a and is always between the geometric mean and the arithmetic mean. This makes this metric a perfect alternative among log-based measures.

4 The TPC-D Single-Stream Metric

The TPC-D benchmark consisted of 17 queries numbered Q1 through Q17 and two refresh functions UF1 and UF2. The single-stream metric is a "query per hour" rate using the inverse of the geometric mean of the query times. However, the TPC subcommittee in charge of developing the benchmark realized that very small query times would pose a problem opted for the following solution: the minimum query elapsed time that can be reported cannot be less than the largest observed query time divided by 1000. The a-displaced geometric mean was considered but not retained because it used values that were not observed ("measured" plus the displacement would be actually used instead of the "measured" values).

Very soon after the TPC-D benchmark became official, vendors started noticing that it was more advantageous to get better performance from the small queries. As a result, to get a good number one would need to have a lot of very small elapsed times in order to boost the power number based on the geometric mean. In order to increase TPC-D power stream results, a number of techniques were developed such as semantic query optimization, materialized join structures and finally aggregated single table and aggregated join structures. The use of materialized aggregated structures eventually made the benchmark useless and it was retired in 1999 leading to the TPC-H benchmark that prevents the use of materialized structures. Even though the "minimum query time" rule is still in vigor in the TPC-H benchmark it never applied since the inception of the benchmark.

Both the a-displaced average and the TPC-D single-stream metric are geometric-based and provide undue advantage to very small query times although they avoid the main pitfall. The TPC-D metric gives extra incentive to lowering the largest query but either one would be very harmful in the following case. Let us consider a real-life case of an optimization of TPC-D benchmark query 1, a full scan of the lineitem table with a large aggregation. With an easily defined pre-aggregated structure the query is reduced to scan of the structure. At scale factor 100 the size of the table is about 600 million rows while the structure is only a few thousand rows. As a result, the elapsed time goes from minutes to less than a second while the updating of the small structure has virtually no impact on the inserts (UF1) and deletes (UF2). To illustrate the impact of the approach assume for the sake of argument that all query times are 100 seconds. With query 1 time going to 0.2 second the arithmetic mean goes from 100 second to 95 seconds – a 5% improvement - while the geometric mean goes from 100 seconds to 72 seconds – a 28% improvement. Had the arithmetic mean been used, the hyper-inflation in single-stream metric may not have occurred and it is even possible that the relative rankings of the results would have been the same – this point is made in a different context in [5].

5 The Case for the Arithmetic Mean

In addition to its hyper-sensitivity to small values the geometric mean is difficult to "sell" especially in the context of a decision-support benchmark. The first difficulty is its relatively complicated formula. But the main problem is that it does not relate to a physical quantity that can be readily understood by users. In the context of decision-support, elapsed times are what a system is measured against. In a single-stream context where a number of queries are run back to back only the total elapsed time and the average query elapsed time have physical meaning. Elapsed times are absolute numbers and the only operations that make sense for users in this context are additions. In the sequel we show that under these conditions the only valid metric is the arithmetic mean.

In [2] the authors demonstrated that the geometric mean was the only valid metric to summarize normalized numbers. Using a similar argument we will show here that the only valid metric to summarize single-stream elapsed times in the context of a decision-support benchmark is the arithmetic mean. We will first establish properties that such a metric $f(x_1,x_2,..,x_n)$ should have.

If all observations are equal to some value a then the metric itself must be equal to a. In other words, whatever the value of a

$$f(a,a,...,a) = a \tag{11}$$

Since all queries must be treated equally then any permutation of the values x_1, x_2,..,x_n must provide the same value, i.e. for all permutations a_{i1}, a_{i2}, .. a_{in}.

$$f(a_{i1}, a_{i2}, .., a_{in}) = f(a_1, a_2, .., a_n) \tag{12}$$

Finally, we want the metric to have meaning in the context of absolute elapsed times i.e. we want an additive property. Indeed, if we were to run the same benchmark on two machines, then, to aggregate the results we should be able to add the individual metrics obtained on the individual machines, i.e.

$$f(a_1 + b_1, a_2 + b_2, ..., a_n + b_n) = f(a_1, a_2, .., a_n) + f(b_1, b_2, ..., b_n) \tag{13}$$

Using the properties above it is very easy to see that

$$a = f(a,a,...,a) = f(a,0,..,0) + f(0,a,0,..,0) + ... + f(0,0,...,a) \tag{14}$$

Hence

$$a = nf(a,0,0,...) \text{ and } f(a,0,..,0) = \frac{a}{n} \tag{15}$$

Similarly

$$\begin{gathered} f(0,b,0,..,0) = \frac{b}{n} \text{ and} \\ f(a,b,...) = f(a,0,...,0) + f(0,b,0,...0) + ... = \frac{a}{n} + \frac{b}{n} + .. \end{gathered} \tag{16}$$

and consequently, the metric is equal to the arithmetic mean.

6 Conclusion

In this paper we have summarized the arguments for and against the arithmetic mean and the geometric mean. We have also provided esoteric metrics similar to the geometric mean – some new - but we made the case for simplicity and meaning. In the context of decision-support we have shown through the example of the TPC-D single stream metric that a choice made a priori due to the nature of industry standard benchmarks led to unexpected results. Finally, we hope to have shown that the best way to handle a metric for a decision-support benchmark is the arithmetic mean.

References

1. TPC BENCHMARK D (Decision Support), Transaction Processing Council, http://www.tpc.org
2. Fleming, P., Wallace, J.: How Not to Lie With Statistics: The Correct Way to Summarize Benchmarks. Comm. ACM 29(3), 218–221 (1986)
3. Licea-Kane, B.: http://www.spec.org/gwpg/gpc.static/geometric.html
4. Crolotte, A.: Issues in Metric Selection and the TPC-D Single Stream Power, http://www.tpc.org
5. Smith, J.: Characterizing Computer Performance with a Single Number. Communications of of the ACM 31(10) (October 1988)
6. Kurian John, L.: More on Finding a Single Number to indicate Overall Performance of a Benchmark Suite. ACM SIGARCH Computer Architecture News 32 (March 2004)
7. Citron, D., Hurani, D., Gnadrey, A.: The Harmonic or Geometric Mean: Does it Really Matter? ACM SIGARCH Computer Architecture News 34(4) (September 2006)
8. Mashey, J.: War of the benchmark Means: Time for a Truce. ACM SIGARCH Computer Architecture News 32(4) (September 2004)
9. Calot, G.: Cours de Statistique Descriptive, Dunod, Paris (1964)