Set Processing vs Row Processing

Author: Craig Shallahamer (craig@orapub.com), Version 1j, 24-May-2012.

Background and Purpose

The purpose of this notepad is to see if set based processing provides increased throughput and scalability compared to row based processing.

The set based processing was performed using SQL. The row based processing was performed using Oracle's plsql.

Experimental Data

Below is the some of the the experimental data. It only incluces the sample times (elapsed time) to process 100K rows. (Not 200K, ... 1000K rows.) The experiment was run on a Dell single six-core CPU, Oracle 11.2G. According to "cat /proc/version": Linux version 2.6.32-300.3.1.el6uek.x86_64 (mockbuild@ca-build44.us.oracle.com) (gcc version 4.4.4 20100726 (Red Hat 4.4.4-13) (GCC)) #1 SMP Fri Dec 9 18:57:35 EST 2011.

```
setProcessing = {.068537, .065366, .064761, .064817,
   .065704, .065924, .065158, .065038, .065117, .065771, .065061, .065039};
rowProcessing = {242.97254, 243.032936, 244.847836, 245.510885, 245.722446, 244.850288,
   247.626378, 245.650096, 244.134109, 245.05695, 244.900598, 243.624751};
```

Basic Statistics

In this section I calculate the basic statistics, such as the mean and median. My objective is to ensure the data has been collected and entered correctly and also to compare the two datasets to see if they appear to be different.

```
myData = {
   {"Set", N[Mean[setProcessing]], N[StandardDeviation[setProcessing]],
    Length[setProcessing], DistributionFitTest[setProcessing]},
   {"Row", N[Mean[rowProcessing]], N[StandardDeviation[rowProcessing]],
    Length[rowProcessing], DistributionFitTest[rowProcessing]}
  };
toGrid = Prepend[myData, {"Processing\nType", "Avg Time (s)", "Stdev", "Samples", "P-Value"}];
Grid[toGrid, Frame → All]
```

Processing Type	Avg Time (s)	Stdev	Samples	P-Value
Set	0.0655244	0.00101909	12	0.00133837
Row	244.827	1.29425	12	0.548369

Sample Set Normality Tests

Before we can perform a standard t-test hypothesis tests on our data, we need to ensure it is normally distributed...because that is one of the underlying assumptions and requirements for properly performing a t-test.

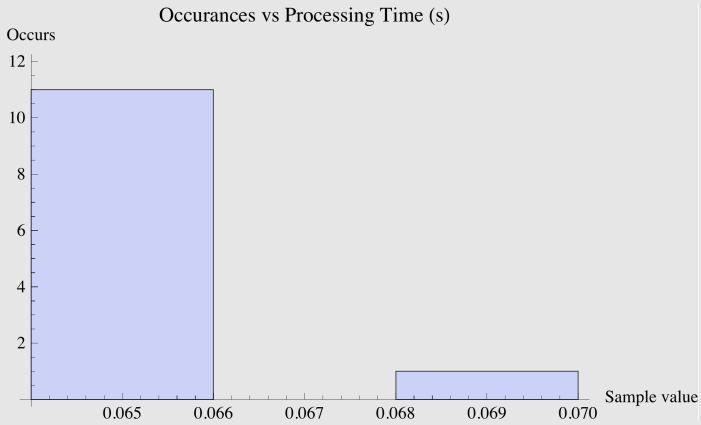
Statistical and vIsual normality test

Our alpha will be 0.05, so if the distribution fit test results in a value greater than 0.05 then we can assume the data set is indeed normally distributed

The first test is just to double check to make sure my thinking is correct. Since I creating a normal distribution based on a mean and standard deviation (just happens to be based on the my sample set data), I would expect a p-value (the result) to greatly exceed 0.05. Notice that the more samples I have created (the final number), the closer the p-value approaches 1.0.

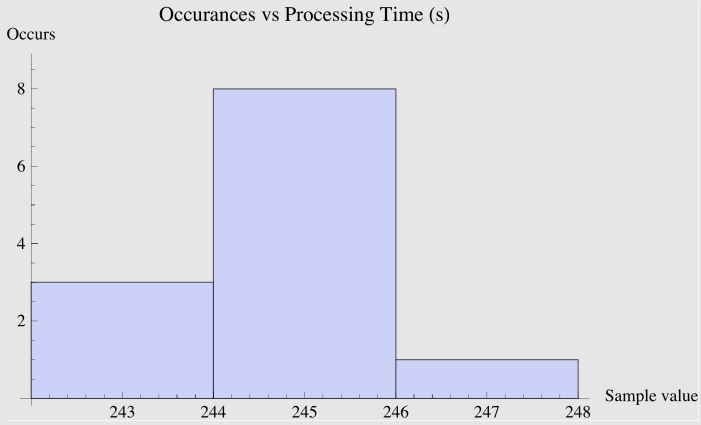
```
DistributionFitTest[setProcessing]
Histogram[setProcessing, PlotLabel → "Occurances vs Processing Time (s)",
 AxesLabel → {"Sample value", "Occurs"}]
```

```
0.00133837
```




```
DistributionFitTest[rowProcessing]
Histogram[rowProcessing, PlotLabel → "Occurances vs Processing Time (s)",
 AxesLabel → {"Sample value", "Occurs"}]
```

```
0.548369
```



Sample Comparison Tests (when normality exists)

Assuming our samples **are normally distributed**, now it's time to see if they are significantly different. If so, then we know changing the commit write optoins indeed makes a significant performance difference...at least statistically.

The null hypothesis is; there is no real difference between our samples sets. We need to statistically prove that any difference is the result of randomness; like we just happened to pick poor set of samples and it makes their difference look much worse than it really is.

A t-test will produce a statistic p. The p value is a probability, with a value ranging from zero to one. It is the answer to this question: If the populations really have the same mean overall, what is the probability that random sampling would lead to a difference between sample means larger than observed?

For example, if the p value is 0.03 we can say a random sampling from identical populations would lead to a difference smaller than you observed in 97% of the experiments and larger than you observed in 3% of the experiments.

Said another way, suppose I have a single sample set and I copy it, resultling in two identical sample sets. Now suppose we perform a significance test on these two identical sample sets. The resuting p-value will be 1.0 because they are exactly the same. We are essentially doing the same thing here except we have to different sample sets... but we still want to see if they "like" each other..and in our case we hope they are NOT like each other, which means the p-value will low... below our cut off value of 0.05.

For our analysis we choose alpha of 0.05. To accept that our two samples are statistically similar the p value would need to be less than 0.05 (our alpha).

Good reference about the P-Value and significance testing: http://www.graphpad.com/articles/pvalue.htm

Here we go (assuming our samples are normally distributed):

1. Our P value threshold is 0.05, which is our alpha.
2. The null hypothesis is the two populations have the same mean. (Remember we have to sample sets, which not the population.)
3. Do the statistical test to compute the P value.
4. Compare the result P value to our threshold alpha value. If the P value is less then our threshold, we will reject the null hypothesis and say the difference between our samples is significant. However, if the P value is greater than the threshold, we cannot reject the null hypothesis and any difference between our samples are not statistically significant.

```
TTest[{rowProcessing, setProcessing}]
```

TTest::nortst : At least one of the p-values in {0.548369, 0.00133837}, resulting from a test for normality, is below 0.025`. The tests in {T} require that the data is normally distributed. ≫

1.31671×10^{-26}

If the above T-Test results (p value) are less then our threshold we can say there is a significant difference between the two sample sets.

Sample Comparison Tests (when normality may NOT exist)

If our sample sets are **not normally distributed,** we can not perform a simple t-test. We can perform what are called location tests. I did some research on significance testing when non-normal distributions exists. I found a very nice reference:

http://www.statsoft.com/textbook/nonparametric-statistics

The paragraph below (which is from the reference above) is a key reference to what we're doing here:

...the need is evident for statistical procedures that enable us to process data of "low quality," from small samples, on variables about which nothing is known (concerning their distribution). Specifically, nonparametric methods were developed to be used in cases when the researcher knows nothing about the parameters of the variable of interest in the population (hence the name nonparametric). In more technical terms, nonparametric methods do not rely on the estimation of parameters (such as the mean or the standard deviation) describing the distribution of the variable of interest in the population. Therefore, these methods are also sometimes (and more appropriately) called parameter-free methods or distribution-free methods.

Being that I'm not a statistician but still need to determine if these sample sets are significant different, I let *Mathematica* determine the appropriate test. Notice that one of the above mentioned tests will probably be the test *Mathematica* chooses.

Note: If we run our normally distributed data through this analysis (speically, the "LocationEquivalenceTest"), *Mathematica* should detect this and use a more appropriate significant test, like a t-test.
Here we go with the hypothesis testing (assuming our sample sets are not normally distributed):

2. The null hypotheses is the two populations have the same mean. (Remember we have to sample sets, which is not the population.)
3. Do the statistical test to compute the P value.
4. Compare the result P value to our threshold alpha value. If the P value is less then our threshold, we will reject the null hypothesis and say the difference between our samples is significant. (Which is what I'm hoping to see.) However, if the P value is greater than the threshold, we cannot reject the null hypothesis and any difference between our samples are not statistically significant; randomness, picked the "wrong" samples, etc.

```
LocationEquivalenceTest[{rowProcessing, setProcessing}, {"TestDataTable", "AutomaticTest"}]
SmoothHistogram[{rowProcessing, setProcessing}]
Histogram[{rowProcessing, setProcessing}]
```

	Statistic	P-Value
Kruskal-Wallis	17.28	4.30948×10^{-8}

, KruskalWallis