



How “Normal” Is Your IT Data?

TECHNICAL WHITE PAPER

Table of Contents

Abstract..... 3
Introduction 4
Data Description 6
Results.....7
Conclusions 10
VMware’s Solution..... 10
References 10
Appendix A..... 11
Metric: OS Metric - Free Memory11
Metric: Application Metric – IIS Total Not Found Errors..... 12
Metric: User Metric – Number of hits.....13
Metric: Quality of Service Metric – Login Response Time15

Abstract

This study attempts to look at a variety of IT performance data (operating system metrics, quality-of-service metrics, business performance metrics, etc.) to determine how well these metrics can be modeled as “normally distributed.” This study came about as a result of multiple IT management vendors providing metric baselining (or dynamic thresholding) capability based on the assumption of normally distributed data. These solutions compute means and standard deviations of the metric data and a set number of standard deviations are used to determine the dynamic baselines for these metrics.

This study will show that IT data does not behave normally and in fact there are an infinite number of distributions such data can take as the distributions change with respect to time. Thus any techniques using data mean and standard deviations as the basis of their analysis for real-time baselining will be significantly inaccurate.

Introduction

The use of statistical analysis in IT has been a hallmark of capacity planning for some time. The techniques developed for such analysis has made effective use of parametric modeling for certain types of data sets. However, more and more IT systems management vendors are introducing the same types of analysis for real time IT problem identification. While using such modeling techniques on long range aggregated data looking for capacity issues may be useful, the usage of the same techniques for real time, short range, non-aggregated data for problem determination may not. The most prevalent technique being employed by systems management vendors is the assumption of normal distribution for every metric analyzed. In this paper various types of IT data are examined for how close (or far) they are from being normally distributed. By having a means of measuring the degree of closeness to a normal distribution, one can better ascertain the effectiveness of such techniques.

One other topic to be introduced is the notion of parametric vs. non-parametric analysis. Parametric techniques presume a distribution (or a set of distributions) for the data set and then based on the known behavior of those distributions one can quickly arrive at potential behavioral patterns of the data. In contrast, Non-Parametric techniques make no presumptions regarding the behavior of the data and use ranking and counting methods for similar computations. The advantages of parametric techniques are if the data set is close to the presumed distribution then this type of analysis can provide an accurate and fast computation of data behavior. However, as will be shown later in this paper, non-aggregated IT data in general does not conform to any type of distribution thus requiring the usage of non-parametric methods for real-time behavior determination.

Normal distribution is the typical bell-shaped curve taught in many introductory mathematics and statistical courses. These types of bell shaped curves often arise as a result of random processes (such as measurement errors, or population sampling, etc.). The mean and standard deviations can be obtained using:

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$$

$$\sigma = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2}$$

where x_i are the time series values. To put an upper and lower threshold on a time series data that is normally distributed, a set number of standard deviations from the mean can be used as a measure of the desired fraction of data to include within the upper and lower thresholds. For example, to include 95% of the data within the band, two standard deviations from the mean will provide the appropriate upper and lower boundaries. This is depicted in the image below.

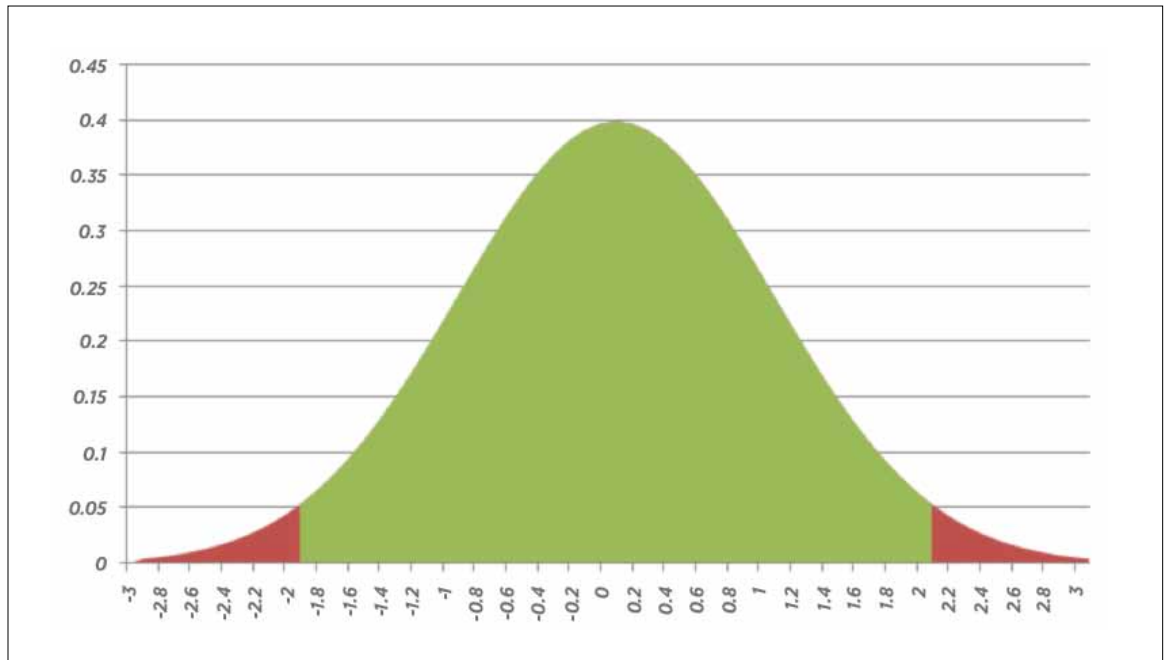


Figure 1. Typical Normal Distribution

This has become a popular method of determining bands of normalcy within IT data as it provides an easy and computationally efficient mechanism for determining the upper and lower thresholds.

One of the methods currently employed for providing dynamic thresholding is to segment the data into presumed cycles (weekday-weekend cycles for example) and then using the segmented data as the metrics for the computation of the dynamic threshold for that specified cycle. For example, assuming that there is a cyclical pattern to the data such that each hour for every day of the week behaves differently, we can segment the time series data into 7 day segments and each of those segments can be broken into 24 hour segments. To determine the dynamic threshold for say, Monday at 9 AM, all data for this daily/hourly segment is assembled and, using the assumption of normal distribution, an upper and a lower threshold is computed as described above. The behavior of data once segmented can differ from the overall behavior and thus the analysis provided in the following sections will also attempt to cover the behavior characteristics of segmented data.

Data that is nearly normal can be analyzed in the fashion outlined above. To determine whether a data set is nearly normal a statistical test needs to be performed. In this study, the Kolmogorov-Smirnov (K-S) test was utilized to test for normalcy. This test involves the following steps:

- 1 Obtain the cumulative probability distribution of the actual data set and the theoretical normal distribution.
- 2 Determine the maximum difference between the distributions above. This is the test statistic for the K-S test.
- 3 Using the K-S distribution, obtain the critical values for the test statistic for the desired significance level.
- 4 If the test statistic is less than the critical value, then the normal distribution assumption is valid within the desired significance level.

The details of this test can be found in the cited references at the end of this paper. This test is used throughout this paper to determine whether a specific IT metric is close enough to be normally distributed.

The first step in the above process required obtaining the actual distribution of the data. Average Shifted Histogram (ASH) technique is used to obtain the distribution. This technique allows one to very closely model the data distribution without having to rely upon typical histogram limitations of wide band data buckets (see image below).

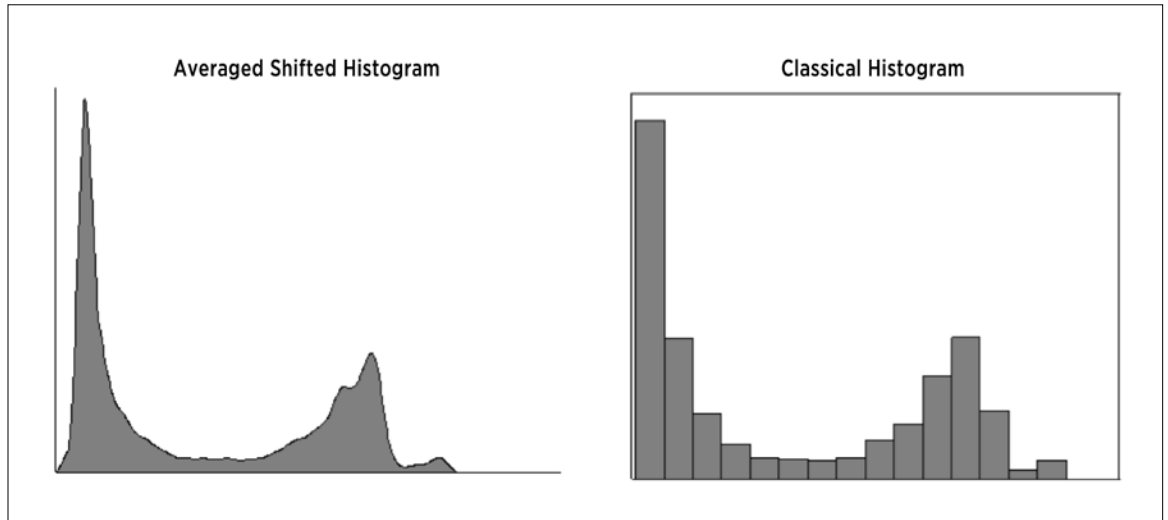


Figure 2. Resolution of ASH vs. Classical Histogram Distribution Representation

The ASH works by starting with the traditional banding of data into N buckets each of size h , but then shifting each of these N buckets $M-1$ times to average the behavior across the data set using:

$$P_h(x) = \frac{1}{M} \sum_{l=0}^{M-1} \frac{1}{Nh} \sum_{i=1}^N \left\{ \sum_j I(X_i \in V_{jl}) I(x \in V_{jl}) \right\}$$

$$V_{jl} = [(j + l/M)h, (j + 1 + l/M)h], \quad l = 0, \dots, M - 1; \quad j = 0, \dots, N - 1$$

where $P_h(x)$ is the distribution of the data set and $I(X_i \in V_{jl})$ represents the count of number of x_i which are elements of the set V_{jl} . More information regarding this technique can be obtained from the cited references.

Data Description

The data set for this study consisted of real world IT and business data regarding a variety of verticals, from a Customer Service application, to a Stock Trade application, to Ad-Serving. The data sets consisted of operating system, quality of service (user experience monitoring), business, and environmental metrics (temperature, fan speed, etc.). These data sets represent the spectrum of potential metrics within IT used to find clues to root cause of problems and thus require statistical analysis for determination of abnormality. The data sets are summarized in Table 1.

APPLICATION	TECHNOLOGIES	MONITORING TOOLS
Ad-Serving Application	Linux servers, Apache, MySql	Nagios, Ganglia, Cacti
Transaction Settlement Application	Solaris Servers, Oracle DB	Quest Foglight
Customer Service Application	Windows servers, Oracle Clusters	HP Sitescope, RUM, BPM
Stock Trade Application	Linux Servers, Custom Application	HP Openview, SNMP

Table 1. Summary of Data Sets

The variety of data collection mechanisms and the diversity of the applications ensured a wide enough sampling to enable a significant confidence in the results and conclusions reached.

Results

The results obtained from the above data sets will be presented on a per application basis. The table below displays the value of the test statistic (D) for the various data segments of the measured metrics. If the test statistic is less than a value of 0.038 (critical value of K-S statistics at the 1 percent criticality level) then the data can be assumed to be normally distributed. For all results the significance level chosen for the K-S test is 1 percent, which is much more lenient in accepting normal distribution. In typical circumstances a significance value of 5 percent is more customary.

The further away the test statistic is from the critical value, the further the distribution is from being normal. Test statistic values above 0.1 represent a significant departure from normal distribution. The data for each metric was also segmented into three one hour slots (9 AM on Monday, Tuesday, and Wednesday) for the normalcy test. In the following tables the %Normal column represents the percentage of data segments which satisfied the normalcy criteria and the Average D column represents the average value of the test statistics for the specified metric segments.

SUMMARY BY	TOTAL SEGMENTS	AVERAGE D	%NORMAL
App Metrics	35	.178	0.0%
OS Metrics	91	.243	0.0%
Bus. Metrics	4	.097	0.0%
All Data	34	.208	0.0%
Monday 9AM	30	.207	0.0%
Tuesday 9AM	30	.226	0.0%
Wednesday 9AM	36	.241	0.0%
All Metrics	130	.221	0.0%

Table 2. Ad-Serving Application

SUMMARY BY	TOTAL SEGMENTS	AVERAGE D	%NORMAL
App Metrics	65	.284	0.0%
OS Metrics	53	.188	1.80%
QoS Metrics	11	.290	0.0%
All Data	35	.247	0.0%
Monday 9AM	32	.259	0.0%
Tuesday 9AM	32	.239	0.0%
Wednesday 9AM	30	.233	0.0%
All Metrics	129	.245	0.78%

Table 3. Transaction Settlement Application

SUMMARY BY	TOTAL SEGMENTS	AVERAGE D	%NORMAL
OS Metrics	87	.223	2.30%
App Metrics	34	.158	2.94%
User Metrics	48	.134	6.25%
QoS Metrics	12	.252	0.0%
All Data	43	.177	4.65%
Monday 9AM	46	.197	2.17%
Tuesday 9AM	46	.197	4.35%
Wednesday 9AM	46	.184	2.17%
All Metrics	181	.189	3.32%

Table 4. Customer Service Application

SUMMARY BY	TOTAL SEGMENTS	AVERAGE D	%NORMAL
Env. Metrics	59	.310	0.0%
OS Metrics	51	.249	0.0%
All Data	25	.235	0.0%
Monday 9AM	28	.297	0.0%
Tuesday 9AM	28	.320	0.0%
Wednesday 9AM	29	.282	0.0%
All Metrics	110	.282	0.0%

Table 5. Stock Trade Application

These results clearly show that the behavior of IT data, across a variety of collection sources and data types, does not resemble normal distribution. The average value of the test statistic (D) also shows the metric distributions are very far from being normal (values of D greater than 0.1). When looking at the actual distributions of the metrics we find a variety of distributions, some that can be characterized, and some that cannot. For example the image below shows the actual distribution of memory usage and clearly indicates a multinomial distribution which could be characterized had the information been available a priori.

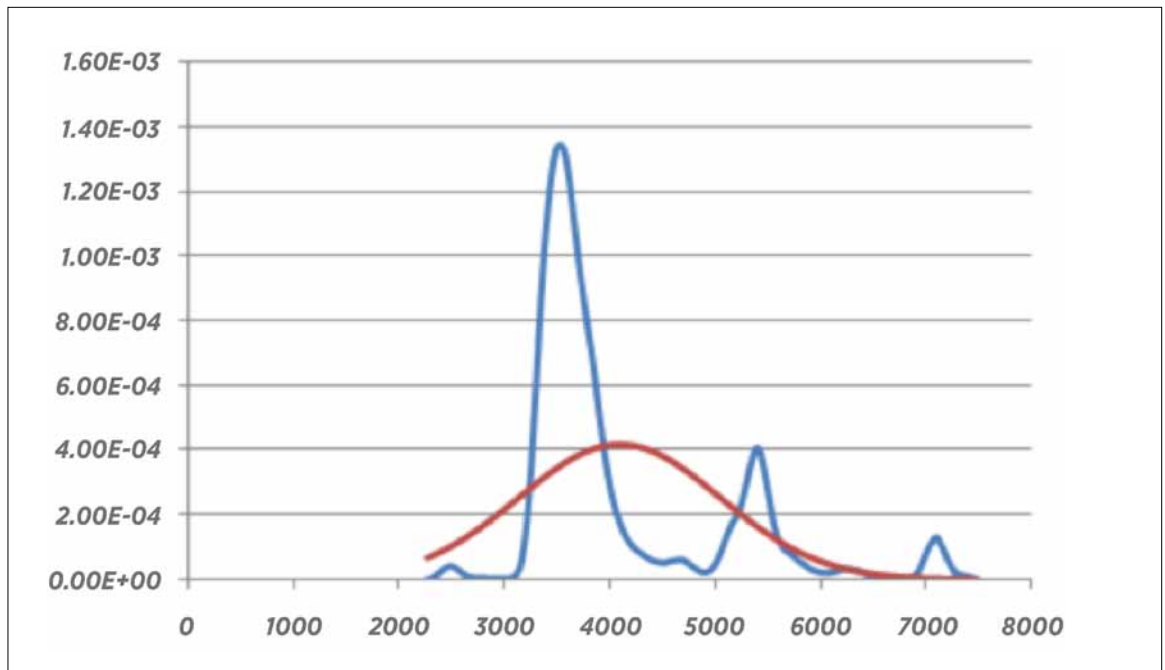


Figure 3. Distribution of Memory Usage

The same metric looked at in a time window of Tuesdays at 9AM shows behavior which cannot be characterized:

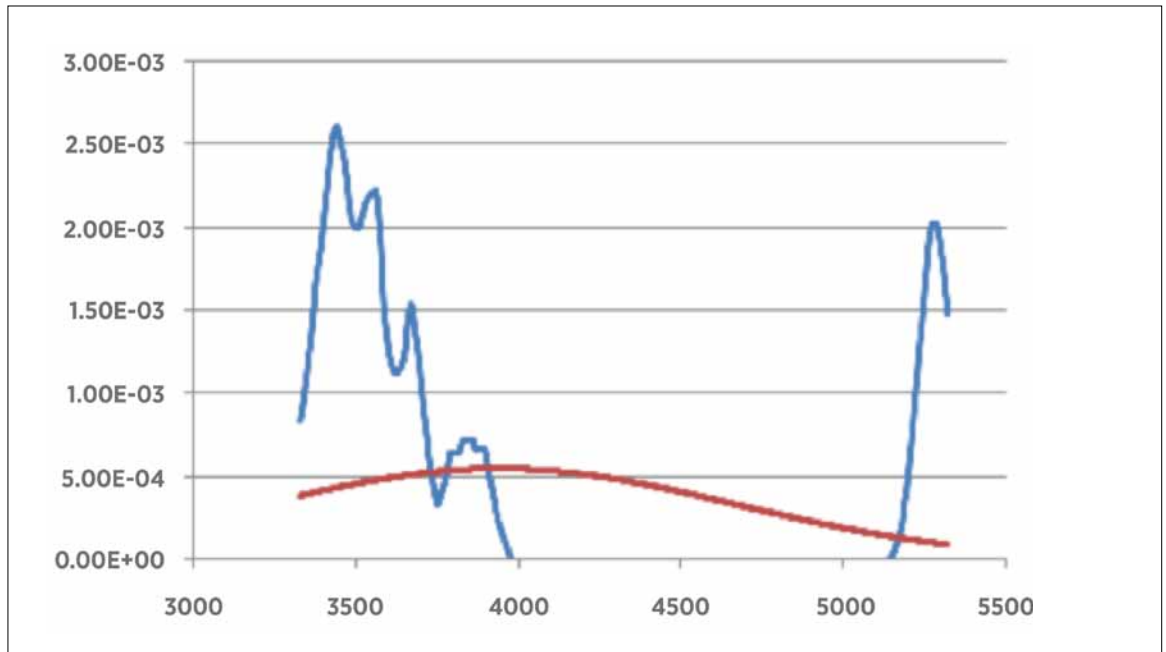


Figure 4. Distribution of Memory Usage for Tuesday 9am Window

With this kind of data it is not possible to use parametric methods for analysis as clearly there is no way to characterize the distributions in a generic way. Appendix A shows more examples of metrics and their distributions.

Conclusions

From the data presented above it is quite clear that IT data does not resemble normally distributed data. In fact, there are so many varieties of distributions that it renders any parametric analysis ineffective. In the rare cases where the data does come close to being normally distributed there does not seem to be any discernable pattern in the type or behavior profile of the metric to use it in a consistent way. Because of this overwhelming evidence of non-conformity to presumed distributions, any methodology employing a simplistic normal distribution assumption or any other parametric analysis must be questioned as to their validity.

VMware’s Solution

The recognition of the ineffectiveness of parametric methods for determining dynamic thresholds has led VMware to develop a set of non-parametric methods for data analysis. The other key insight of our extensive research is that no single algorithm can be effective in analyzing the myriad of data types present in IT environments. The VMware vCenter™ Operations product incorporates eight different non-parametric techniques for determining the best thresholds for IT data. And since the algorithms are data agnostic they can work with Boolean type data (up/down metrics), batch (sparse data sets), and even text based metrics to determine whether they are acting normally or abnormally. The key to the accuracy of the algorithms is not only in their non-parametric approach, but also the fact that the exact cycles of data are uncovered and appropriately utilized for best threshold determination. All of these features combine to provide the most comprehensive and accurate determination of thresholds for the entire enterprise.

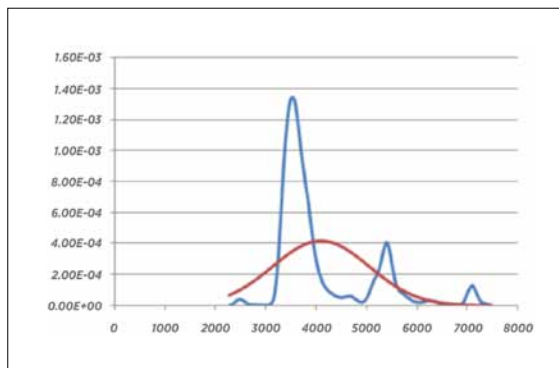
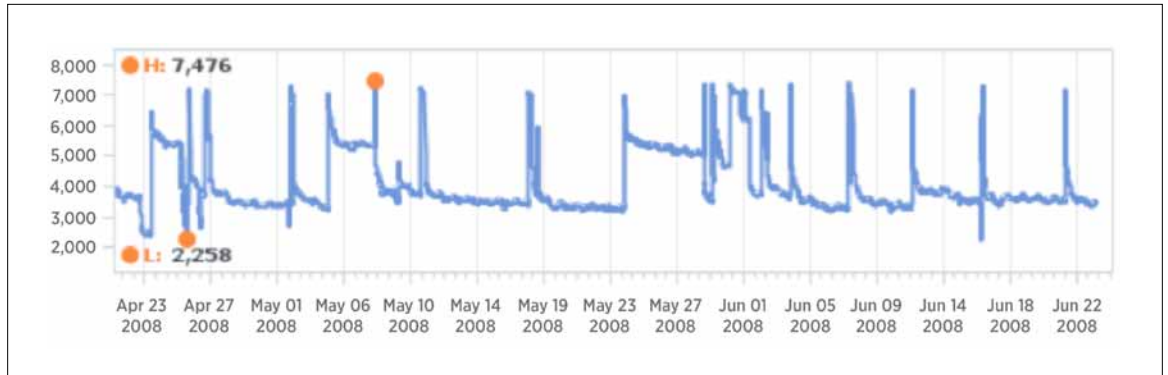
References

- Porkess, R.: *Web-linked Dictionary Statistics*, HarperCollins, 2006
- Jaynes, E.T.: *Probability Theory The Logic Of Science*, Cambridge, 2003
- Spiegel, M., Schiller J., Srinivasan R., *Probability and Statistics*, McGraw-Hill, 2000

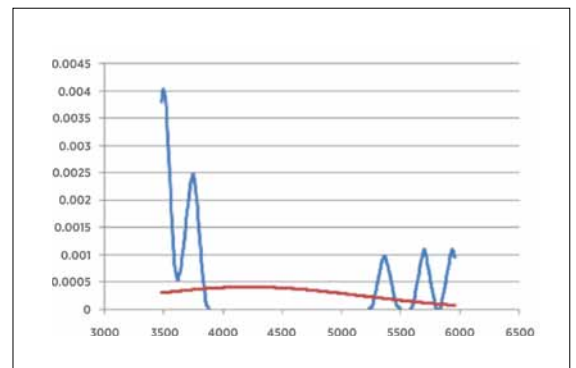
Appendix A

This appendix shows various metrics and their actual distributions (blue) and the idealized normal distribution (red)

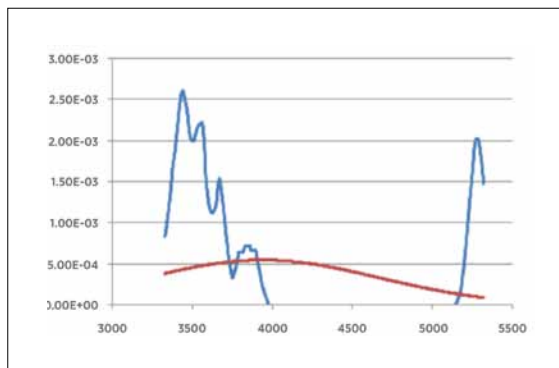
Metric: OS Metric - Free Memory



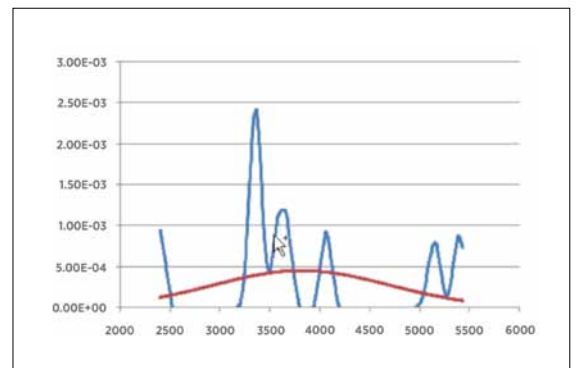
Overall Distribution



Monday 9 AM Distribution



Tuesday 9 AM Distribution



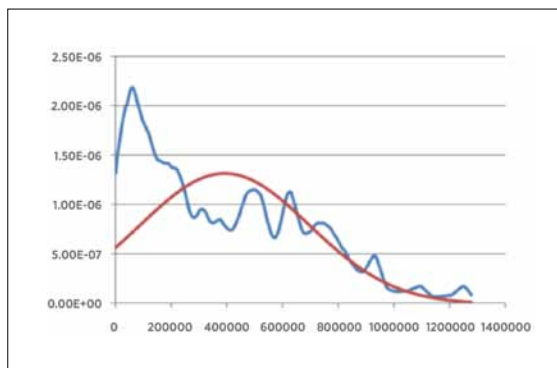
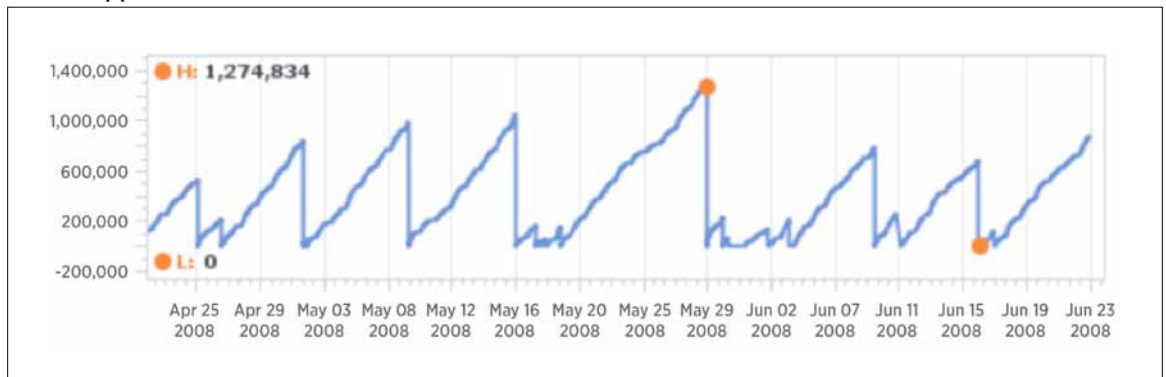
Wednesday 9 AM Distribution

K-S test results:

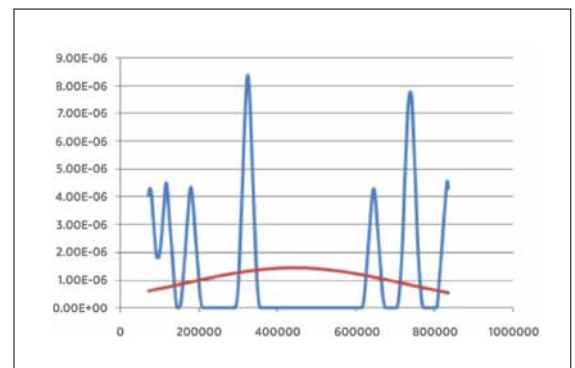
DATA SET	TEST STATISTIC	CRITICAL VALUE	PASS/FAIL NORMALCY
All data	0.246	0.0384	Failed
Monday 9 AM	0.332	0.0615	Failed
Tuesday 9 AM	0.308	0.0615	Failed
Wednesday 9AM	0.193	0.0615	Failed

Comments: Metrics such as memory typically behave in a multimodal fashion (as is evident from the above actual distributions). The Normal distribution assumption can be quite poor in such circumstances as the data typically congregates at the tail of the idealized normal distribution for such data. From the K-S test results we can see the wide variability between the different days at 9 AM (from the value of the test statistic) and the significant departure from the critical value thus leading to a failure of Normalcy test.

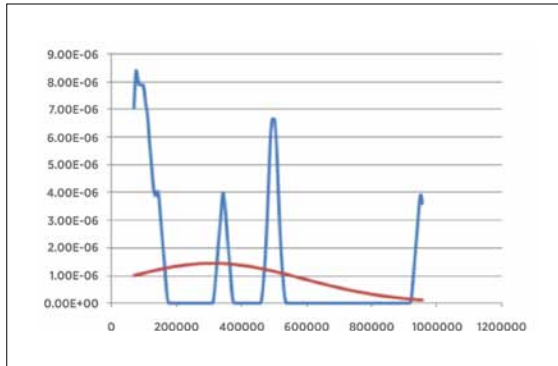
Metric: Application Metric – IIS Total Not Found Errors



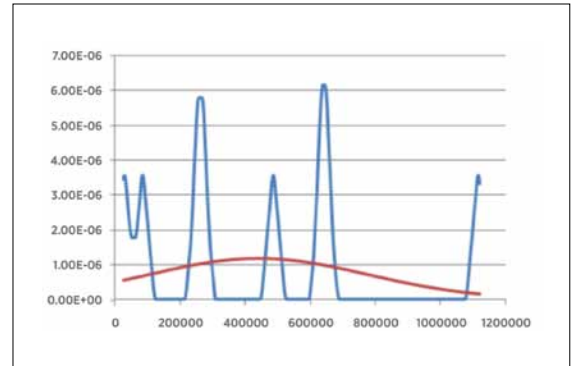
Overall Distribution



Monday 9 AM Distribution



Tuesday 9 AM Distribution



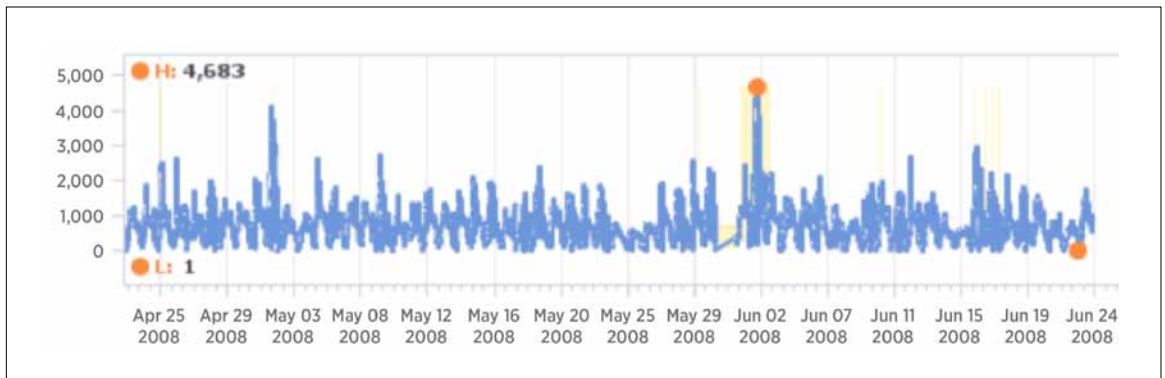
Wednesday 9 AM Distribution

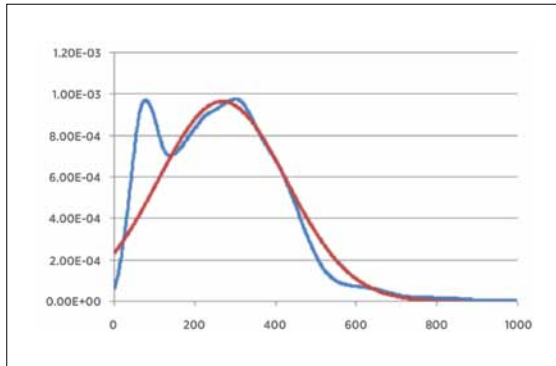
K-S test results:

DATA SET	TEST STATISTIC	CRITICAL VALUE	PASS/FAIL NORMALCY
All data	0.0966	0.0384	Failed
Monday 9 AM	0.193	0.0384	Failed
Tuesday 9 AM	0.238	0.0384	Failed
Wednesday 9AM	0.159	0.0384	Failed

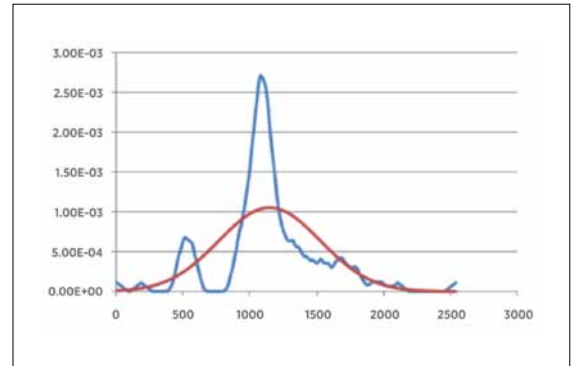
Comments: Application type metrics such as those shown above often track the restarts of the application and can be quite difficult to model.

Metric: User Metric – Number of hits

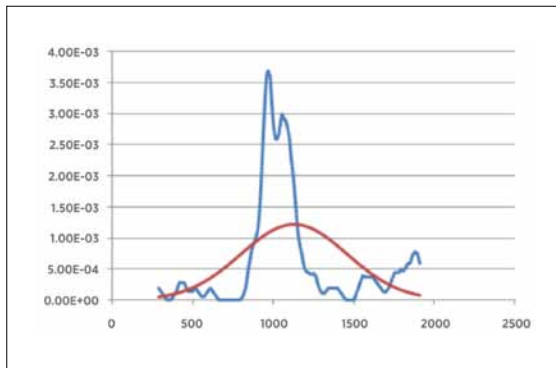




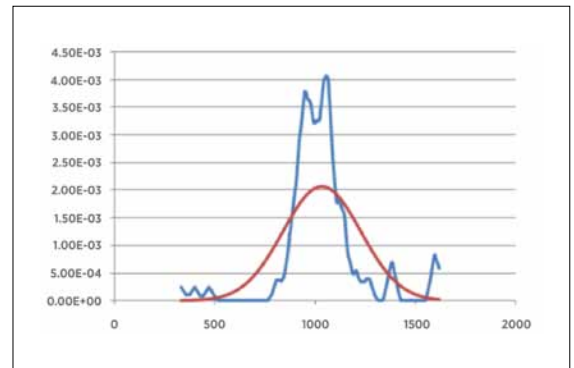
Overall Distribution



Monday 9 AM Distribution



Tuesday 9 AM Distribution



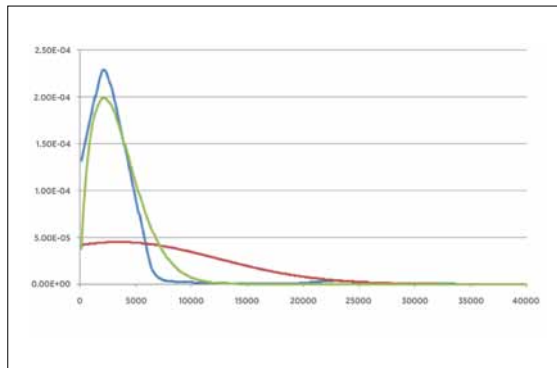
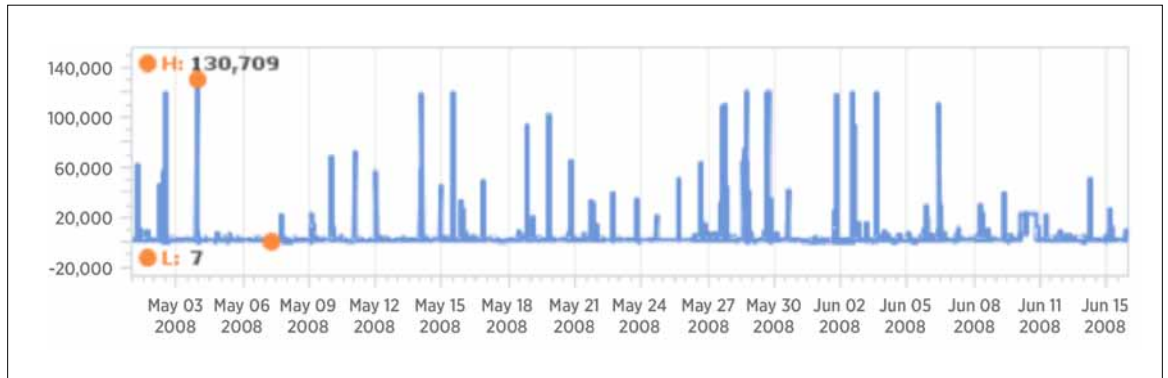
Wednesday 9 AM Distribution

K-S test results:

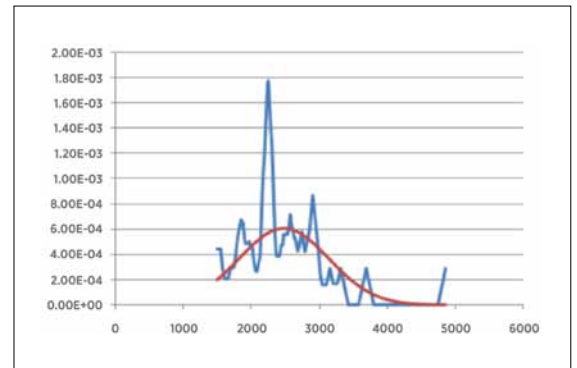
DATA SET	TEST STATISTIC	CRITICAL VALUE	PASS/FAIL NORMALCY
All data	0.0525	0.0384	Failed
Monday 9 AM	0.126	0.0384	Failed
Tuesday 9 AM	0.221	0.0384	Failed
Wednesday 9AM	0.142	0.0384	Failed

Comments: User behavior metrics (such as hits, time spent, download size, etc.) often exhibit bimodal behavior as shown in the overall distribution above. Although the overall distribution looks very close to a normal distribution (as enumerated by the K-S test statistic value), once broken down into hours of the day and days of the week, the data starts to rapidly deviate from normal distribution.

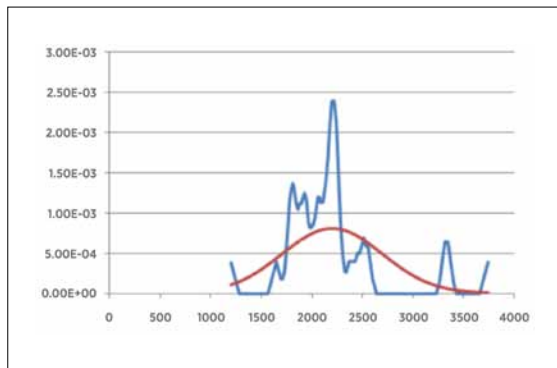
Metric: Quality of Service Metric – Login Response Time



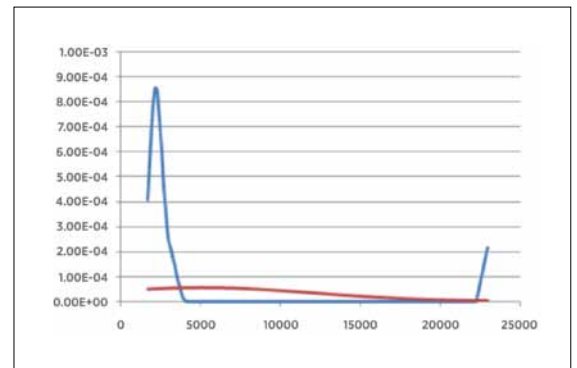
Overall Distribution



Monday 9 AM Distribution



Tuesday 9 AM Distribution



Wednesday 9 AM Distribution

K-S test results:

DATA SET	TEST STATISTIC	CRITICAL VALUE	PASS/FAIL NORMALCY
All data	0.343	0.0384	Failed
Monday 9 AM	0.095	0.0384	Failed
Tuesday 9 AM	0.199	0.0384	Failed
Wednesday 9AM	0.491	0.0384	Failed

Comments: Response time metrics typically exhibit the overall behavior shown above, which may be approximately modeled as a Weibull distribution as shown in overall distribution graph (green). However, the segmented data can vary greatly from day to day and has no resemblance to any known distributions, thus leading to ineffective parametric modeling.



VMware, Inc. 3401 Hillview Avenue Palo Alto CA 94304 USA Tel 877-486-9273 Fax 650-427-5001 www.vmware.com

Copyright © 2011 VMware, Inc. All rights reserved. This product is protected by U.S. and international copyright and intellectual property laws. VMware products are covered by one or more patents listed at <http://www.vmware.com/go/patents>. VMware is a registered trademark or trademark of VMware, Inc. in the United States and/or other jurisdictions. All other marks and names mentioned herein may be trademarks of their respective companies. Item No: VMW-TECH-WP-HOW-NORMAL-IS-YOUR-DATA-USLET-101