Software Performance Analytics in the Cloud

Kingsum Chow and Wanyi Zhu
Alibaba Inc
Hangzhou, China
kingsum.chow@gmail.com, wanyi.wyz@alibaba-inc.com

ABSTRACT

The emergence of large-scale software deployments in the cloud has led to several challenges: (1) measuring software performance in the data center, and (2) optimizing software for resource management. This tutorial addresses the two challenges by bringing the knowledge of software performance monitoring in the data center to the world of applying performance analytics. It introduces data transformations for software performance metrics. The transformations enable effective applications of analytics. This tutorial starts with software performance in the small and ends with applying analytics to software performance in the large. In software performance in the small, it summarizes performance tools, data collection and manual analysis. Then it describes monitoring tools that are helpful in performance analysis in the large. The tutorial will guide the audience in applying analytics to performance data obtained by common tools. This tutorial describes how to select analytical methods and what precautions should be taken to get effective results.

CCS Concepts

• Computer systems organization □ Architectures □ Distributed architectures • Software and its engineering □ Software organization and properties □ Extra-functional properties □ Software performance.

Keywords

Datacenter efficiency; software performance; capacity planning; analytics.

1. INTRODUCTION

The tutorial will guide the audience in applying analytics to performance data obtained by using mostly freely downloadable tools. This tutorial describes how to select analytical methods and what precautions should be taken to get effective results. The following is the outline of the tutorial.

(1) Software Performance Analytics: Past, Present and Future
(2) Software Performance 101
(3) Introduction to Analytics
(4) Putting it together: Applying analytics to software performance in the cloud and data centers

We started with a brief overview of the classic text on Introduction to Algorithms [9]. Then we moved on to a set of common systems performance tools [12]. To put them together on computer servers running nowadays, we applied knowledge gain from the classic text on Computer Architecture: A Quantitative Approach [11]. We put together the knowledge of software performance analysis [4]. We also coached performance engineers 0 in using a set of mostly freely available tools and a massive amount of data that need to be analyzed [6][8] together on a modern computer server running multiple processes. We applied design of experiments [7] to aid computer performance analysis. We used a collection of scripts to process and analyze the data. Among them, R becomes a cornerstone of our analysis process. R is a great statistics tool. But to deal with the vastly different computer performance data sources, we found Advanced R [10] comes in handy in bring analytics together. As this tutorial is targeting software performance engineers, this paper will cover some details of analytics. The full tutorial will cover details of putting software performance and analytics together.

2. ANALYTICS

Statistical techniques have been applied in many areas of computer performance measurement, such as designing experiments and simulation, analyzing, predicting and optimizing performance. Owing to the stochastic nature of computer systems, measuring performance is prone to noise and experimental errors. The field of statistics provides a rich set of tools and techniques to effectively deal with such noise and errors, allowing one to obtain meaningful measurements. This section will focus on one key application of statistical analysis the anomaly detection, and two methodologies (univariate and multivariate analysis), which have been deployed in many data center management.

2.1 Univariate Analysis

The selection of anomaly detection methods depends on different data situations such as the distribution, degree of skewness and sample sizes. This section reviews several univariate methods.

3.1.1 Parametric Method. One simple classical approach to screen outlier is to use the Standard Deviation (SD) method. It is defined as any observations beyond two or three SD above and below the mean of the observations may be considered as outliers. This is based on the assumption that the data follows a parametric distribution. According to Chebyshev inequality, if a random variable X with mean u and variance exists, no more than $1/k^2$ of the distribution's values can be more than k standard deviations away from the mean- $P(|x-\mu| >= k\sigma) <= 1/k^2$ .The

ICPE'17, April 22 - 26, 2017, L'Aquila, Italy

ACM 978-1-4503-4404-3/17/04...$15.00
DOI: http://dx.doi.org/10.1145/3030207.3053676

inequality has great utility because it can be applied to any probability distribution in which the mean and variance are defined, and enable us to determine the likelihood of having extreme values in the data. It is limited in that it only gives the smallest proportion of our data within k standard deviations of the mean. For instance, at least 75%, 89% and 94% of the data are within 2, 3, 4 standard deviations of the mean, respectively.

Another similar univariate method that can be used to screen data for outliers is the Z-Score, using the mean and standard deviation. $Z_i = \frac{x_i - \bar{x}}{sd}$ $, where\ x_i \sim N(\mu, \sigma^2)$ and sd is the standard deviation of data. When X follows a normal distribution, Z-scores that exceed 3 in absolute value are generally considered as outliers. It presents a reasonable criterion for identification of the outlier when the sample size is larger than 10. However, this method can be affected by extreme values and cause a masking problem, i.e., the less extreme outlier go undetected because of the most extreme outliers, and vice versa. To avoid this problem, the median and the median of the absolute deviation of the median (MAD) are employed in modified Z-Score ($M_i = \frac{0.6745\,(x_i - \tilde{x})}{MAD}$) where $\mathrm{MAD} = \mathrm{median}\{|\,X_i - \tilde{x}|\}, \tilde{x}\ is\ the\ sample\ median.$ The observations are suggested to be labeled as outliers when |Mi| > 3.5 through the simulation based on pseudo-normal observations for sample size larger than 10.

3.1.2 Non-parametric Method – Tukey's univariate boxplot. In parallel with robust statistics, practical methods for analyzing data evolved known as Exploratory Data Analysis (EDA). It is common to consider Tukey's schematic boxplot as an informal test for flagging the existence of outliers. A significant feature of this method is that it does not assume an underlying probability distribution for the data and therefore is flexible in practical settings. The Tukey univariate boxplot is specified by five parameters: the two extremes, the upper Q3 (75th percentile) and lower Q1 (25th percentile) quartiles and median. Data outside the outer fences (Q1 – 3 IQR, Q3 + 3 IQR) are considered to be extreme outliers. Tukey's boxplot method is quite effective for large data set without any prior assumption of data distribution.

2.2 Multivariate Analysis

Even though statistical tests are quite powerful under well-behaving statistical assumptions such as distribution assumption, most distributions of real-world data may be unknown, or may not follow specific distributions, or include correlated features.

The common multivariate anomaly detection has been deployed in many data centers management to offline fixed threshold using training data- such as Multivariate adaptive statistical filtering (MASF) maintain a separate threshold for data segmented and aggregated by time [1]. With the increasing complexity of performance data, real-time detection have been proposed as opposed to traditional methods that rely on static profiling or limited sets of historical data.

These online algorithms can be categorized into three groups. 1) Time series analysis. This becomes quite popular for continuous performance monitoring system used widely in cloud data centers, such as monitoring spikes in user engagement on the data platform, and detecting anomalies in system metrics after a new software release [2]. 2) Statistical learning: This includes a combination of Tukey method and the multinomial goodness-of-fit test using the Relative Entropy statistic. This approach has demonstrated sufficient accuracy and effectiveness for monitoring and detecting the 'anomaly' in performance data from production environment to data captured from multi-tier web applications running on server class machines [3]. 3) Cluster analysis: Outliers can be detected using distance/density measures where objects that are at a substantial distance from any other cluster are considered as outliers. Many clustering algorithms- KMeans, Hierarchical clustering algorithm, and Density-Based Clustering have been utilized in anomaly detection of cloud data center.

With the data center environment increasing in scale and complexity, these machine-learning algorithms enable automating the detection of suspicious anomalies in a responsive and accurate manner.

ACKNOWLEDGMENTS

We thank multiple teams in Alibaba that has developed multiple data collectors, storage and analysis portals for multiple projects.

REFERENCES

[1] J.P. Buzen and A.W. Shum, "MASF: multivariate adaptive statistical filtering," in Int. Computer Measurement Group (CMG) Conf., Nashville, TN, USA, Dec. 4-8, pp. 1-10, 1995.

[2] Owen Vallis , Jordan Hochenbaum , Arun Kejariwal, A novel technique for long-term anomaly detection in the cloud, Proceedings of the 6th USENIX conference on Hot Topics in Cloud Computing, p.15-15, June 17-18, 2014, Philadelphia, PA

[3] Chengwei Wang, Krishnamurthy Viswanathan, Lakshminarayan Choudur, Vanish Talwar, Wade Satterfield, and Karsten Schwan. 2011. Statistical techniques for online anomaly detection in data centers. In Proceedings of the IFIP/IEEE International Symposium on Integrated Network Management (IM'11). IEEE, 385--392.

[4] Kingsum Chow, Pooja Jain and Khun Ban, "Tutorial: Java Application Performance in the Data Center, Data Collection and Analysis" CMG imPACt 2016 Conference. November 7-10, 2016 in La Jolla, California.

[5] Kingsum Chow, Li Chen and Colin Cunningham, "How We Coach Performance Engineers to Adopt Data Science in Reproducible Analytics", Intel Analytics Summit, March 22-24, 2016, Santa Clara, California.

[6] Kingsum Chow and Pranita Maldikar "Applying Analytics to Workload Optimized Systems" [best paper award] presented at the FastPath workshop, held in conjunction with the IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS), March 23-25, 2014.

[7] Keerthi Palanivel, Kingsum Chow, Khun Ban and David Lilja, "A Stepwise Approach to Software-Hardware Performance Co-Optimization Using Design of Experiments", Computer Measurement Group Performance and Capacity 2013, Nov 5th to 7th, La Jolla, California, USA.

[8] Shruthi Deshpande, Kingsum Chow, Peng-fei Chuang and Latifur Khan, "Big Data Analysis to Characterize Workload using Machine Learning Algorithms for High Dimensional Performance Data", Computer Measurement Group Performance and Capacity 2013, Nov 5th to 7th, La Jolla, California, USA.

[9] Cormen, Thomas H., Charles Eric. Leiserson, and Ronald L. Rivest. *Introduction to Algorithms DC*. Cambridge, MA: MIT, 1989. Print.

[10] Wickham, Hadley. *Advanced R*. Boca Raton, Fla.: Chapman & Hall, 2015. Print.

[11] Hennessy, John L., David A. Patterson, and Krste Asanovicì. *Computer Architecture: A Quantitative Approach*. Amsterdam: Elsevier, 2012. Print.

[12] Gregg, Brendan. *Systems Performance: Enterprise and the Cloud*. Upper Saddle River, NJ: Prentice Hall, 2014. Print