Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 94 (2016) 324 – 331

Procedia Computer Science

The 11th International Conference on Future Networks and Communications (FNC 2016)

Performance Evaluation of Techniques to Detect Discontinuity in Large-Scale-Systems

Haroon Malik[a]*, Elhadi M. Shakshuki[b]

[a]*Weisberg Division of Computer Science, Marshall University, WV, USA*
[b]*Jodery School of Computer Science, Acadia, University, NS, Canada*

Abstract

Contemporary data centres rely heavily on forecasts to accurately predict future workload. The accuracy of a forecast greatly depends upon the merit of performance data fed to the underlying algorithms. One of the fundamental problems faced by analysts in preparing data for use in forecasting is the timely identification of data *discontinuities*. A discontinuity is an abrupt change in a time-series pattern of a performance counter that persists but does not recur. We used a supervised and an unsupervised techniques to automatically identify the important performance counters that are likely indicators of discontinuities within performance data. We compared the performance of our approaches by conducting a case study on the performance data obtained from a large scale cloud provider as well as on open source benchmarks systems. The supervised counter selection approach has superior results in terms of unsupervised approach but bears an overhead of manual labelling of the performance data.

Peer-review under responsibility of the Conference Program Chairs

Keywords: Anomaly detection; Performance test; Data center

1. Introduction

As organizations grow, their data centers inevitably become larger and more complex. These centers usually provide numerous services to the internal and external users; such services are commonly deployed on virtual machines (VMs) across a cluster of computing devices, which can either be the property of the organization (i.e., a "private

* Corresponding author. Tel.: +1-304-696-5655.
E-mail address: malikh@marshall.edu

Peer-review under responsibility of the Conference Program Chairs
doi:10.1016/j.procs.2016.08.048

cloud") or be rented from a public cloud provider such as Amazon EC2. In both cases, private and public, managing "the cloud" requires appropriate virtualization and cost-effective provisioning of the infrastructure with respect to the type and size of the service requests (i.e., the workload). The implementation of this virtualization and provisioning needs to be carefully considered. Traditionally, most of the research on hardware provisioning of data centers and service management of cloud environments has been focused on the dynamic allocation methods, load migration, virtual machine consolidation, and balancing strategies.

Performance of all these algorithms depends heavily on the quality of input data (i.e., the predicted/forcasting workload), Hence, practitioners and data scientists spend considerable time (e.g., up to 80%) in preparing data for their forecast algorithms. One of the fundamental problems faced by analysts in preparing data for long-term forecast is the identification and removal of data discontinuities. To date, there does not exist any automated approach to assist data center operators in detecting discontinuities in the performance data.

A discontinuity is an abrupt change in a time-series pattern that persists but does not reoccur. Examples include a) a significant change in a counter's value b) a significant change in the slope (rate of change) of the counter's value, c) a significant change in a cycle or amplitude or both. Discontinuities do not occur instantaneously, but over a brief period called a transition period. If an analyst recognizes that a discontinuity has occurred, they may want to ignore the early data and base their forecast on the measurements taken after the discontinuity. Moreover, detecting a discontinuity provides an analyst a reference point to retrain their forecasting models and make necessary adjustments. Therefore, analysts require automated techniques that can identify discontinuities among thousands of performance counters collected across hundreds of machines.

In our previous work[2], we introduced an unsupervised and supervised approach to reduce the volume of performance counter by synthesizing performance signatures. These performance counters are minimal set of performance counters that describe the essential characteristics of the system being monitors (actively or passively). Furthermore, we used unsupervised approach to detect discontinuity in for identifying discontinuities in performance data from 5,000 machines over a span of seven years[10]. In this paper we compare the performance of both supervised and unsupervised approach in detecting discontinuities in performance data of large scale systems.

2. Overview of the Methodology

In this section of the paper we provide an overview of our methodology for discontinuity detection using both supervised and unsupervised signature techniques.

2.1. Unsupervised Approach

The performance logs obtained from the production environment consists of thousands of performance counters. Many of the performance counters are either invariants such as *'Component Uptime'*, '*Component Last Failure'* or are configuration constants, such as *'No of DB Connections Allowed"*, '*Message Queue Length'* and *'Total Component Memory'*. These counters captures little variance and the values of such performance counters seldom change or correlate to dependent variable such as workload volume. These variables are of little help to analysts in detecting discontinuities. We use a robust and scalable statistical technique i.e., Principal Component Analysis (PCA) to identify a few of the performance counter that capture the maximum variation of the collected data and have the potential to

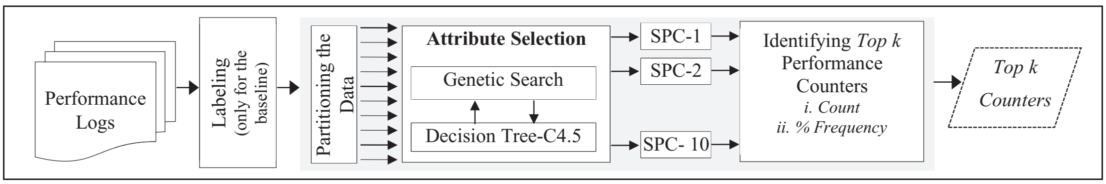

Fig. 1. The steps involved in supervised performance counter selection methodology

capture discontinuities in their time series counter values. We choose PCA due to a) our previous success in using it with performance data of a large-scale system and b) its superior performance in identifying performance counters that are sensitive to minute changes in both workload and environment as compared to many other supervised and