ARTICLE IN PRESS

Future Generation Computer Systems ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Detecting performance anomalies in scientific workflows using hierarchical temporal memory

Maria A. Rodriguez *, Ramamohanarao Kotagiri, Rajkumar Buyya

Cloud Computing and Distributed Systems (CLOUDS) Laboratory, School of Computing and Information Systems, The University of Melbourne, Australia

HIGHLIGHTS

- An anomaly detection framework for scientific workflows is presented.
- HTM is used to detect anomalies on a stream of resource consumption time series data.
- The HTM-based model is unsupervised and learns incrementally.
- The framework is platform-agnostic and can be deployed on different infrastructures.
- Detected anomalies can trigger scheduling and resource provisioning actions.

ARTICLE INFO

Article history:
Received 23 November 2017
Received in revised form 18 April 2018
Accepted 9 May 2018
Available online xxxx

Keywords:
Online anomaly detection
Scientific workflow
Hierarchical temporal memory
Performance anomalies

ABSTRACT

Technological advances and the emergence of the Internet of Things have lead to the collection of vast amounts of scientific data from increasingly powerful scientific instruments and a growing number of distributed sensors. This has not only exacerbated the significance of the analyses performed by scientific applications but has also increased their complexity and scale. Hence, emerging extreme-scale scientific workflows are becoming widespread and so is the need to efficiently automate their deployment on a variety of platforms such as high performance computers, dedicated clusters, and cloud environments. Performance anomalies can considerably affect the execution of these applications. They may be caused by different factors including failures and resource contention and they may lead to undesired circumstances such as lengthy delays in the workflow runtime or unnecessary costs in cloud environments. As a result, it is essential for modern workflow management systems to enable the early detection of this type of anomalies, to identify their cause, and to formulate and execute actions to mitigate their effects. In this work, we propose the use of Hierarchical Temporal Memory (HTM) to detect performance anomalies on real-time infrastructure metrics collected by continuously monitoring the resource consumption of executing workflow tasks. The framework is capable of processing a stream of measurements in an online and unsupervised manner and is successful in adapting to changes in the underlying statistics of the data. This allows it to be easily deployed on a variety of infrastructure platforms without the need of previously collecting data and training a model. We evaluate our approach by using two real scientific workflows deployed in Microsoft Azure's cloud infrastructure. Our experiment results demonstrate the ability of our model to accurately capture performance anomalies on different resource consumption metrics caused by a variety of competing workloads introduced into the system. A performance comparison of HTM to other online anomaly detection algorithms is also presented, demonstrating the suitability of the chosen algorithm for the problem presented in this work.

1. Introduction

Scientific applications enable the extraction of knowledge from vast amounts of data. The volume and underlying value of these data are continuously increasing as technological advances that support the creation of more powerful and precise scientific instruments are made. The significance and importance of the analyses performed by these applications becomes then of utmost importance for scientific progress. For instance, the Large Hadron Collider (LHC) at CERN produces approximately 30 petabytes of data per year that must be analyzed to understand the effects of particle collisions. Another example is the Laser Interferometer Gravitational Observatory (LIGO) project, which harnesses scientific workflows to process data. Since the deployment of their advanced interferometers leveraging hardware developments in optics and

* Corresponding author.
E-mail addresses: marodriguez@unimelb.edu.au (M.A. Rodriguez), kotagiri@unimelb.edu.au (R. Kotagiri), rbuyya@unimelb.edu.au (R. Buyya).

https://doi.org/10.1016/j.future.2018.05.014
0167-739X/

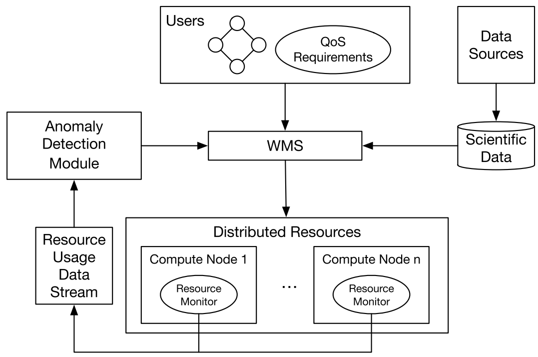

Fig. 1. High-level overview of a performance anomaly detection framework for scientific workflows.

vibration suspension systems, gravitational waves stemming from the collision of black holes were detected for the first time in 2015. This discovery is not only fundamental in supporting Einstein's General Theory of Relativity, but has also paved the way for the emerging field of gravitational wave astronomy which will eventually enable a better understanding of gravitational wave sources and the universe.

This increased ability to collect data from different sources leads to an inevitable growth in the scale and complexity of scientific applications. This is also true for workflows, which have been traditionally used as a way of structuring different scientific computations and their dependencies. They enable scientist to describe applications in a platform agnostic manner while Workflow Management Systems (WMS) hide the complexities of the underlying computing infrastructure by transparently orchestrating the execution of workflow tasks. Emerging extreme-scale workflows are becoming more widespread and so is the need to efficiently automate their execution on a variety of platforms such as high performance computers, dedicated clusters, cloud computing infrastructures, and, more recently, fog environments.

It becomes paramount then to execute large-scale complex scientific workflows in a reliable and scalable manner on distributed systems. In particular, being able to continuously monitor their performance and create models of expected behavior that adapt over time to the dynamicity of the underlying infrastructure can be of great benefit for scientists. Specifically, we argue that WMSs can use runtime resource consumption information to detect performance anomalies and mitigate their effects on the overall workflow execution and Quality-of-Service (QoS) requirements. This can be achieved by making dynamic scheduling and resource provisioning decisions that are triggered by the detection of such anomalous behavior. For example, a performance anomaly may be detected in the CPU consumption pattern of a running task. This will eventually lead to the task taking longer than expected to complete, ultimately having a negative impact on the total workflow execution time (i.e., makespan). Such effect becomes more prominent as the scale of the applications in terms of the number of tasks, the amount of data they process, and their computational requirements continue to grow. Resource management modules can attempt to correct performance issues by provisioning more resources, executing unscheduled workflow tasks on more powerful resources, migrating or replicating running tasks, or scaling the compute resources vertically (e.g., increasing the memory of a virtual machine), among other approaches.

As a result, our goal is to develop a framework that optimizes the performance of complex data-intensive workflows by detecting, diagnosing, and potentially correcting the cause of anomalous runtime performance. An overview of the key components of such a system are depicted in Fig. 1. In this work, we focus on the first challenge; that is, the early detection of performance anomalies through real-time infrastructure monitoring. In particular, we aim to identify patterns in the data that do not conform to expected behavior and we focus on analyzing time series data that contain the resource consumption details of tasks at different stages of their execution. The metrics considered in our approach are related to the CPU and I/O usage of a given task on a particular machine.

There are various requirements that arise from the nature of the problem addressed and the monitored data. Mainly, we strive to implement a model that is capable of learning in an online fashion as data becomes available. The main reason for this requirement is the nature of the computing infrastructures used for the deployment of workflows. Firstly, there are a variety of platforms currently used for this purpose; by having a model that learns as data is collected, our framework can be deployed on different computing infrastructures in a seamless manner. Secondly, there are a wide range of factors impacting the performance of running jobs in compute nodes, especially in increasingly popular multi tenant, virtualized environments such as cloud and fog platforms. By learning incrementally in an online manner, the model will dynamically adjust to environmental changes such as peak hour in a data center. Finally, online learning enables our framework to be used to detect anomalies in the execution of different scientific workflows, without the need of previously training a specific model based on data collected from a dedicated infrastructure for example. Another important requirement is for the framework to be capable of processing sequential streaming data in one pass and as the data becomes available. Finally, we are interested in detecting temporal anomalies on these data, that is, identifying a set of abnormal transitions between patterns as opposed to identifying a single data point that deviates from what is standard (i.e., spatial anomaly).

To achieve these requirements, we propose an anomaly detection model that uses Hierarchical Temporal Memory (HTM) networks. HTMs are continuous learning systems that meet our requirements by mimicking the anatomy of the mammalian neocortex and the behavior of neurons to perform learning, inference, and prediction [1]. They are efficient, tolerant to noise, capable of adapting to changes in the statistics of the data, and capable of detecting subtle temporal anomalies [2]. Our detection algorithm uses an HTM model for each monitored metric to detect CPU or I/O related anomalies. For each workflow task deployed, its resource consumption is monitored at given intervals and analyzed as soon as it becomes available. Our comprehensive evaluation demonstrates the efficiency of our method in identifying anomalies caused by different types of competing workloads on two scientific workflows from the bioinformatics field. Our solution has also been designed in such a way that it facilitates the identification of the cause of the anomalies in our future work.

The rest of this paper is organized as follows. Section 2 presents the related work followed by an overview of HTM systems in Section 3. Section 4 explains the proposed anomaly detection method and Section 5 presents the experimental setup and the evaluation of our solution. Finally, conclusions and future work are outlined in Section 6.

2. Related work

Anomaly detection in sequential data has been extensively researched. There are a variety of existing approaches that are offline and are designed to process data once a model capable of making predictions has been built based on some training data. Netflix's Robust Anomaly Detection (RAD) [3] framework is an example. It is a statistical approach based on Robust Principle Component Analysis (RPCA) [4] that relies on data having high cardinality. Another example is HOT SAX [5], an algorithm capable of finding time

Please cite this article in press as: M.A. Rodriguez, et al., Detecting performance anomalies in scientific workflows using hierarchical temporal memory, Future Generation Computer Systems (2018), https://doi.org/10.1016/j.future.2018.05.014.