Expert Systems with Applications 42 (2015) 980–989

Contents lists available at ScienceDirect

Expert Systems with Applications

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

# Intelligent failure prediction models for scientific workflows

Anju Bala *, Inderveer Chana

*Computer Science and Engineering Department, Thapar University, Patiala, India*

ARTICLE INFO

*Article history:*
Available online 21 September 2014

*Keywords:*
Cloud Computing
Workflows
Failure prediction
Scientific workflows
Machine learning

ABSTRACT

The ever-growing demand and heterogeneity of Cloud Computing is garnering popularity with scientific communities to utilize the services of Cloud for executing large scale scientific applications in the form of set of tasks known as Workflows. As scientific workflows stipulate a process or computation to be executed in the form of data flow and task dependencies that allow users to simply articulate multi-step computational and complex tasks. Hence, proactive fault tolerance is required for the execution of scientific workflows. To reduce the failure effect of workflow tasks on the Cloud resources during execution, task failures can be intelligently predicted by proactively analyzing the data of multiple scientific workflows using the state of the art of machine learning approaches for failure prediction. Therefore, this paper makes an effort to focus on the research problem of designing an intelligent task failure prediction models for facilitating proactive fault tolerance by predicting task failures for Scientific Workflow applications. Firstly, failure prediction models have been implemented through machine learning approaches using evaluated performance metrics and also demonstrates the maximum prediction accuracy for Naive Bayes. Then, the proposed failure models have also been validated using Pegasus and Amazon EC2 by comparing actual task failures with predicted task failures.



## 1. Introduction

Cloud Computing has transformed the Information and Communication Technology industry by facilitating on-demand services, elasticity, flexibility and provisioning of computing resources based on utility (Beloglazov & Buyya, 2013). Cloud Computing adopts virtualization technologies to provide various services to the user such as Infrastructure as a Service (IaaS), Hardware as a Service (HaaS), Platform as a Service (PaaS), Software as a Service (SaaS) and Workflow as a Service (WFaaS) (Cushing, Koulouzis, Belloum, & Bubak, 2014; Wang, Korambath, Altintas, Davis, & Crawl, 2014; Zhao, Melliar, & Moser, 2010). The layered architecture of the Cloud services and required tools for these services along with their challenges has been revealed in Fig. 1. As the Cloud offers WFaaS and SaaS on the top layer of Cloud where the user can utilize the services on internet without buying the proprietary rights of software and applications such as Workflow Management System (WMS), scientific applications, E-mail, business and multimedia applications. Some of the key challenges that need to be addressed at software layer are scheduling, data heterogeneity, fault tolerance, fault prediction, interoperability and security etc. The second layer of the architecture acts as a core middleware between software layer and infrastructure layer that endowed PaaS for testing, deploying and controlling web applications which comprise platforms such as Google Apps Engine, Microsoft Azure, Hadoop, Aneka and heroku etc. (Buyya, Shin Yeo, Venugopal, Broberg, & Ivona, 2009). The issues which need to be resolved at this layer are management of big data applications, data analytics and intelligence etc. The infrastructure layer at the bottom consists of IaaS and HaaS that offers various hardware resources and infrastructures on demand without purchasing. Amazon EC2, Eucalyptus, OpenNebula, Nimbus and Open Stack, Rackspace are the examples of infrastructure and hardware providers. Henceforth, Cloud Computing has high availability of these services and thus we are evaluating the use of Cloud services for deploying scientific applications.

Scientific applications are represented as workflows that consist of few tasks to million of tasks which have the dependency between them (Ramakrishnan, Reutiman, Chandra, & Weissman, 2013). The workflow applications for real world business process are identified as Cloud Workflows and WFaaS has been used by some of the researchers for deploying and executing these workflows in Cloud environment. Although, Tan et al. (2009) have also enhanced the performance of real world tumor analysis using the Workflow-as-a-Service in Grid Computing yet they have not defined any architecture for WFaaS in Cloud. Then, Pathirage, Perera, S,

* Corresponding author.
*E-mail addresses:* anjubala@thapar.edu (A. Bala), inderveer@thapar.edu (I. Chana).

http://dx.doi.org/10.1016/j.eswa.2014.09.014
0957-4174/

Kumara, and Weerawarana (2011) proposed architecture for WFaaS to host workflow applications securely in the Cloud but they have not considered Workflow Scheduling. Furthermore, Wang et al. (2014) has defined WFaaS architecture for scheduling the workflow applications to increase the scalability and extensibility. Cushing et al. (2014) have also proposed WFaaS approach for task framing of scientific applications in Cloud. As the work of Juve and Deelman (2010) evaluated Cloud infrastructures as an execution platform for deploying scientific workflows as WFaaS due to the benefits of using Cloud such as provisioning on demand, elasticity, provenance, reproducibility etc. (Deelman, Livny, Berriman, & Good, 2008; Pandey, Karunamoorthy, & Buyya, 2011) have also concluded various advantages of executing scientific workflows with Cloud infrastructure such as cost effective, scalable, decreased runtime, on-demand resource provisioning, and ease of resource management etc. Although, WFaaS is used by only few of the researchers in Cloud, however there are some open challenges that needs to be resolved such as efficient and scalable management of workflows, handling application and resource failures, heterogeneity of data, failure prediction of tasks and fault tolerant scheduling etc. (Deelman, 2009; Gil1 & Deelman, 2007). Among these research issues, the key challenge is to handle the resource and task failures through intelligent prediction of failures for scientific workflows which is not implemented by any of the authors till now.

As most of the existing works employed several of fault tolerant techniques for scientific workflows such as replication, checkpointing, job migration, retry, task resubmission etc. (Bala & Chana, 2012; Ganga, Karthik, & Christopher Paul, 2012). Less research has been done to predict and detect task failures intelligently by adopting machine learning approaches for implementing proactive fault tolerance. As the workflows have used for simulation, high energy physics, astronomy and many other scientific applications. Hence, the reliability models for software and hardware failures cannot be simply applied to handle the task failures proactively (Bala & Chana, 2013; Xie, Dai, & Poh, 2004). It is insightful, if the fault tolerant approach is a reactive one and might not be able to handle the failures intelligently (Varghese, McKee, & Alexandrov, 2010). Henceforth, intelligent task failure detection and prediction is mainly challenging for scientific workflow applications which have lot of job and data dependencies such as Montage, Cybershake, Sipht, Inspiral, Epigenomics and Broadband. These workflows have used in scientific community for various applications such as Montage workflow for astronomical physics, Cybershake workflow for earthquake hazards, Sipht workflow for bioinformatics, Inspiral workflows for detecting gravitational waves, Epigenome for genome sequence operations and Broadband to simulate the impact of an earthquake.

Thus, the goal of proposed models is to predict the task failures intelligently using machine learning approaches before failure occurrence during the execution of scientific workflow applications. The task failures can occur due to overutilization of resources, unavailable resources, execution time or execution cost exceeds than threshold value, required libraries are not installed properly, system running out of memory or disk space and so on. In the present paper, task failures have been generated due to overutilization of resources such as CPU, RAM, Disk Storage and Network Bandwidth. The historical data of task failure parameters has been gathered for training and testing the prediction model in Weka by running multiple scientific workflow applications at different intervals in WorkflowSim. Then, the results of failure prediction model using evaluation metrics have also been compared using machine learning algorithms such as Naive Bayes, Random Forest, LR and ANN and evaluated that the Naive Bayes would be the best machine learning approach for task failure prediction of multiple scientific workflow applications. To validate the accuracy of proposed model, actual failures have been compared with predicted failures using Amazon EC2 and Pegasus. Finally, the experimental results of proposed model have also been compared with existing model after implementing Broadband, Epigenome and Montage.

### 1.1. Motivation for the work

- The motivation of implementing task failure prediction illustrates their inspiration from research challenges of scientific workflows, applications and benefits of using Cloud services for these workflows.
- Our work aspires at analyzing the problem of failure prediction so that Cloud systems would be capable of making autonomic fault tolerant decisions by predicting the task failures with various resource utilization parameters such as CPU utilization, RAM, Disk Storage and Bandwidth utilization.
- Most of the existing works have implemented fault prediction using statistical approaches that would not be useful for predicting failures intelligently, therefore our proposed approach would be useful for predicting task failures proactively for

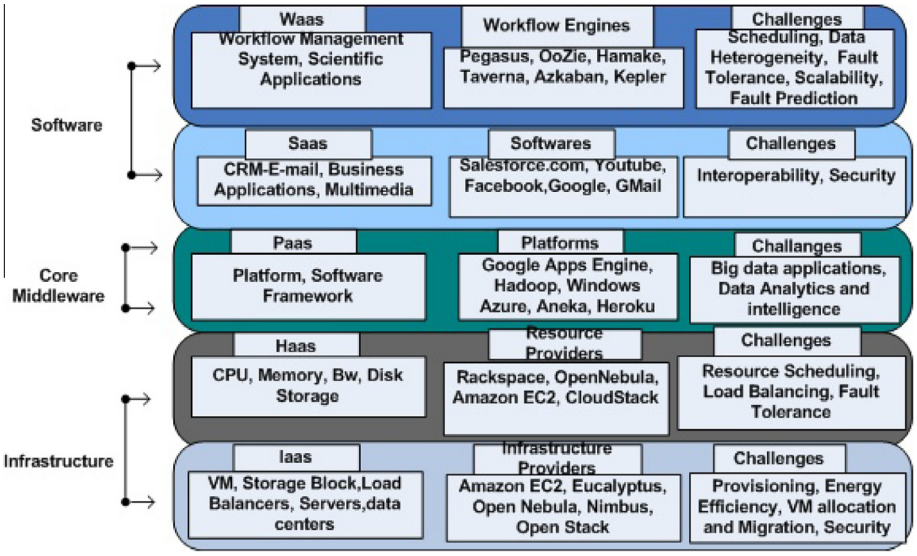


**Fig. 1.** Layered architecture of Cloud along with Waas.