International Journal for Rapid Research in Engineering Technology & Applied Science
Vol. 5 Issue 5 May 2019
ISSN Online: (2455-4723)

A Survey on Data Mining Techniques to Classify Inactive VMs

Vaidehi Bakshi[1], Mohit Jain[2]
1 M.Tech Scholar 2 Head of Department Computer Science
1,2 Department of Computer Science & Engineering
1,2 BM Group of Collage of Engineering And Technology ,Indore, Madhya Pradesh ,India
vaidehibakshi35@gmail.com* , hod.computers@bmcollege.ac.in**

Abstract: **Cloud computing is a new generation computing technology. A wide range of service providers and application developers are accepting the strength of cloud computing. Among various reasons behind efficient computing in cloud the virtualization is one of the essential techniques to obtain the high performance computing. In this context the effective Usage of VMs in a data center is helpful to improve the performance of cloud computing. In this presented work a data mining algorithm is employed for recognizing the inactive VMs in a data center. Therefore a rich survey on data mining techniques is performed. This paper provides the essential contributions in the domain of inactive VM classification. Additionally the key issues and challenges for finding solution are proposed in this paper. The presented solution is implemented in future and the performance improvement is provided.**

Keywords: Cloud Computing, Identifying Inactive VMs, Data Center Management, virtual machine, Supervised Learning

I. Introduction

Cloud computing -Cloud computing is the delivery of on-demand computing services -- from applications to storage and processing power -- typically over the internet and on a pay-as-you-go basis. Cloud computing is the use of various services, such as software development platforms, servers, storage and software, over the internet, often referred to as the "cloud." In general, there are three cloud computing characteristics that are common among all cloud-computing vendors,

1. The back-end of the application (especially hardware) is completely managed by a cloud vendor.

2 .A user only pays for services used (memory, processing time and bandwidth, etc.).

3. Services are scalable

Many cloud computing advancements square measure closely associated with virtualization. The ability to pay on demand and scale quickly is basically a result of cloud computing vendors having the ability to pool resources that will be divided among multiple shoppers. The cloud technology has been gaining great momentum. There square measure a series of layers that square measure interconnected and exist supported the previous layers. Cloud computing has 3 totally different service layers that square measure offered as services. These are:

- **Infrastructure as a Service (IaaS):**
  The first is that the infrastructure layer that's developed on the virtualization technology wherever the service suppliers provide virtual machines as a service to the end-users. It allows IaaS customers to create and discard virtual machines and networks as per their business requirements. They pay for the services they consumed. IaaS removes the necessity for the consumer to invest in procuring and operating physical servers, data storage systems, and other networking resources.
- **Platform as a Service (PaaS):**
  PaaS is the second layer. Here the customers do not manage the virtual servers but rather create the applications within the programming language. They host the programs on the platform services that they obtain. The management and maintenance of the operating systems and other hardware are done by the providers.
- **Software as a Service (SaaS):**
  Applications that don't ought to be put in return below this layer. CRM, email and other office applications come under SaaS. Some services square measure free whereas some square measure beaked monthly or per usage.

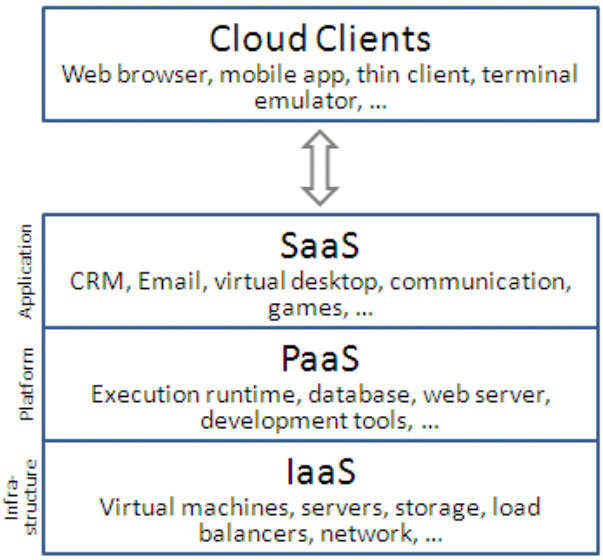

Figure1.cloud services

II. Background:

This section provides the overview of the different terminology that are used during the design and implementation of the proposed system.

Datacenter- Data centers are simply centralized locations where computing and networking equipment is concentrated for the purpose of collecting, storing, processing, distributing or allowing access to large amounts of data. They have existed in one kind or another since the appearance of computers. In the days of the room-sized behemoths that were our early computers, a data center might have had one supercomputer. As instrumentality got smaller and cheaper, and data processing needs began to increase -- and they have increased exponentially -- we started networking multiple servers (the industrial counterparts to our home computers) together to increase processing power. We connect them to communication networks so individuals will access them, or the information on them, remotely. Large numbers of those clustered servers and connected instrumentality is housed in a very area, an entire building or groups of buildings. Today's knowledge center is probably going to own thousands of terribly powerful and really little servers running 24/7.

Supervised learning- Supervised learning, in the context of artificial intelligence (AI)and machine learning, is a type of system in which both input and desired output data are provided. Input and output knowledge area unit labeled for classification to supply a learning basis for future processing. Supervised learning is where you have input variables (x) and an output variable (Y) and you use an algorithm to learn the mapping function from the input to the output.

$$Y = f(X)$$

The goal is to approximate the mapping perform thus well that once you have new computer file (x) that you just will predict the output variables (Y) for that data.

It is known as supervised learning as a result of method of rule learning from the coaching dataset is thought of as a lecturer management the training process. We know the proper answers; the rule iteratively makes predictions on the coaching knowledge and is corrected by the teacher. Learning stops once the rule achieves an appropriate level of performance.

Virtual machine- - A virtual machine (VM) is an operating system (OS) or application environment that is installed on software, which imitates dedicated hardware. The end user has constant expertise on a virtual machine as they'd wear dedicated hardware. Virtual machines (VM) area unit apace replacement physical machine infrastructures for his or her talents to emulate hardware environments, share hardware resources, and utilize a variety of operating systems (OS). VMs provides a additional sturdy security model than ancient machines by providing an extra layer of hardware abstraction and isolation, effective external watching and recording, and on-demand access. However, this new model wants adaptation of existing security ways that, which cannot currently keep up with the ease of creating new VMs with a variety of configurations and lifecycles. Attackers have with success compromised VM infrastructures, permitting them to access different VMs on constant system and even the host. Fortunately, these security concerns are being addressed and users can prevent most intrusions by applying traditional security measures to each VM.

Generic VMs are relatively simple. The hypervisor, additionally referred to as the virtual machine monitor, runs on the host OS and allocates emulated resources to every guest OS. When the guest makes a call the hypervisor intercepts and interprets it into the corresponding call supported by the host OS. The hypervisor controls every VM's access to the mainframe, memory, persistent storage, I/O devices, and therefore the network. Figure 2 shows the architecture for a standard VM infrastructure on a single physical machine.

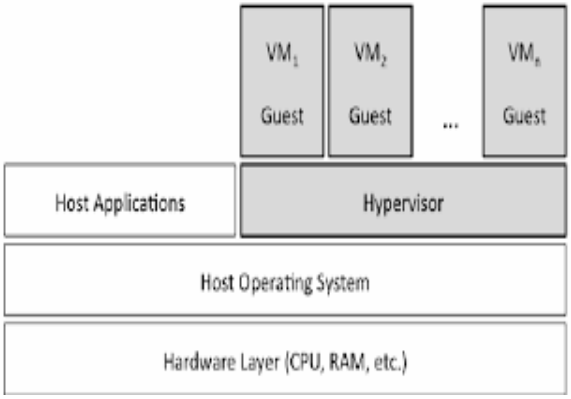

Figure 2 VM Architecture

III. Related Work :

This section includes the different research efforts that are tried to enhance the cloud server performance by using the machine learning approach.

A. Supervised Learning Model for Identifying Inactive VMs in Private Cloud Data Centers

In this paper described a supervised learning model to accurately identify active and inactive VMs for an enterprise private cloud. In order to improve accuracy, we propose an identification model with four steps. The identification model finds fingerprints for inactive VMs first, then determines the purpose of each VM. By leveraging the supervised learning technique (SVM), we identify active VMs with corresponding features to the purpose. Finally, we validate the identification is correct using a network affinity model, and propagate importance to connected VMs. We have demonstrated how we increase the recall of the model. Our evaluation with 750 VMs from enterprise data centers shows the accuracy is 90%, and this is 15% { 20% higher than existing methods. We are closely looking at how we remove the false negatives which can have a significant impact on recommendations to VM users. With the identification results, the current recommendation to VM users is straightforward. If users' VMs are identified as inactive, they are required to take proper actions with our recommendations. i. e. terminating VMs with snapshot, and resizing VMs.

B. ICSI: A Cloud Garbage VM Collector for Addressing Inactive VMs with Machine Learning

To properly detect and address inactive VMs, we present iCSI: a cloud garbage VM collector to improve resource utilization and cost efficiency of enterprise data centers. iCSI includes three main components; a lightweight data collector, a VM identification model, and a recommendation engine. The data collector periodically gathers primitive information from VMs. we design a lightweight approach to periodically gather primitive, but holistic information for running VMs. With the data collection, we describe how we analyze the data in order to extract significant features for active/inactive VM identification. iCSI determines active and inactive VMs through the following steps. iCSI first performs a base case classification for inactive VMs, then determines the purpose of each VM. By leveraging a supervised learning technique (SVM), iCSI identifies active/inactive VMs with corresponding features to the purpose. Finally, iCSI validates the identification results using a network affinity model and propagates the confidence to connected VMs.

C. Virtualization Technology using Virtual Machines for Cloud Computing

In this paper, we tend to gift a system that uses virtualization technology to portion the info center resources dynamically supported the appliance demands and support inexperienced computing by optimizing the number of servers in use. This technique multiplexes virtual to physical resources adaptively supported the ever-changing demand. We use the thought of imbalance metric to mix virtual machines with totally different resource characteristics fitly so the capacities of server's are well utilized. We introduce the thought of "skewness" to live the uneven utilization of the server. By minimizing the imbalance, we will improve the general utilization of the servers within the face of multi-dimensional resource constraints. The thought of inexperienced computing is that the range of physical machines used ought to be reduced as long as they will still satisfy the wants of all virtual machines. Idle physical machines may be turned off to save lots of energy.

D. A Deep Learning based approach to VM behavior identification in cloud systems

Cloud computing information centers square measure growing in size and complexness to the purpose wherever observance and management of the infrastructure become a challenge because of measurability problems. A possible approach to cope with the size of such data centers is to identify VMs exhibiting similar behavior. Existing literature incontestible that agglomeration along VMs that show the same behavior could improve the measurability of each observance and management of an information

center. We propose two deep learning models for the classifier, namely Deep- Conv and DeepFFT based on convolution neural networks and Fast Fourier Transform. We validate our proposal using traces from a real cloud data center and we compare our classifiers with state-of-the-art solutions such as the AGATE technique (that exploits a gray area to adopt the observation time of each VM so that uncertainly classified VMs are not immediately assigned to a group) and a PCA-based clustering solution. The results confirm that the deep learning models consistently outperform every other alternative without the need to introduce a gray area to delay the classification. Even more interesting, the proposed classifiers can provide fast and accurate identification of VMs. In particular, the DeepConv model provides a perfect classification with just 4 samples of data (corresponding to 20 minutes of observation), making our proposal viable also to classify on-demand VMs that are characterized by a very short life span.

IV. Proposed Work :

During the investigation of collected literature a number of research articles are explored additionally a research paper is concluded to extending the given approach the following issues and solutions are proposed for design and implementation.

Problem Formulation – The cloud data centers are providing the efficient computing experience by using the virtualization concept. According to this concept VMs are used for providing the computational efficiency. But when the VMs are not working or inactive for a long term then that is known as the mismanagement of resources. In order to find the inactive VMs in cloud data centers some different approaches are available, among them the machine learning based is also a considerable approach. Due to this technique for inactive VM identification is comparatively new therefore there are scopes to improve the false negative classification rate during the recommendation or prediction. In addition of that for finding the inactive VMs a few features are considered thus it need to find more new features for improving the performance of data centers.

Solution domain -During study and analysis of dataset used in the article [1].The amount of dataset is large in quality or dimensions. In this context the significant resources are required to process the data set. Therefore a suitable feature selection technique is required to train and testing the classifier. Therefore the proposed work leads to design and develop a correlation coefficient based feature selection technique.

In order to rectify the observed issue in the previously made effort in [1] the following improvement is suggested for the proposed research work.

1. **Dataset analysis:** computation of additional significant features for identifying the inactive VM. In addition of that the base features such as Host Information, Resource Usage, Process Information, Network Connections and Login History is also considered for technique development
2. **Classifier improvement:** in this phase that is required to optimize the performance of classification in order to improve the false negative rate of classifier. In the traditional method the SVM (support vector machine) classifier linearly used for classification work. In place of linear configuration of SVM the SVR (support vector regression) model is suggested for implementation and design.

V. Conclusion :

The main aim of the proposed work is to design and develop a supervised learning technique that helps to classify the inactive VMs. The classifications of the inactive VMs are performed on the basis of different VM characteristics or VM features that help to easily recognize the active and inactive VMs. In this context the paper provides the understanding about the cloud computing and their basic components. These components are help to support the cloud computing infrastructure. In addition of the some of the essential and noteworthy contributions are also reported that worked for improving the cloud computing performance by eliminating the inactive VMs. Finally the key issues in the domain of inactive VM classification are reported and the solution for design and development is proposed. In near future the proposed concept is implemented and their performance is reported.

References

[1] M. V. Ahluwalia, A. Gangopadhyay, Z. Chen and Y. Yesha, "Target-Based, Privacy Preserving, and Incremental Association Rule Mining," in *IEEE Transactions on Services Computing*, vol. 10, no. 4, pp. 633-645, 1 July-Aug. 2017.

[2] Amazon Web Services. Data Privacy. https://aws.amazon.com/compliance/data-privacy-faq. ONLINE.

[3] In Kee Kim, Sai Zeng, Christopher Young, Jinho Hwang, and Marty Humphrey. A Supervised Learning Model for Identifying Inactive VMs in Private Cloud Data Centers. In the 17th ACM/IFIP/USENIX Middleware Conference (Middleware '16), Trento, Italy, December 2016.

[4] In Kee Kim, Wei Wang, Yanjun Qi, and Marty Humphrey. Empirical Evaluation of Workload Forecasting Techniques for Predictive Cloud Resource Scaling. In 2016 IEEE International Conference on Cloud Computing (CLOUD '16), San Francisco, CA, USA, Jun. 2016.

[5] Zhiming Shen, Sethuraman Subbiah, Xiaohui Gu, and John Wilkes. CloudScale: Elastic Resource Scaling for Multi-Tenant Cloud Systems. In the ACM Symposium on Cloud Computing (SoCC '11), Cascais, Portugal, October 2011.

[6] Francisco Rocha and Miguel Correia. Lucy in the Sky without Diamonds: Stealing Confidential Data in the Cloud. In the 41st IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W), Hong Kong, China, June 2011.

[7] Liang Zhang, James Litton, Frank Cangialosi, Theophilus Benson, Dave Levin, and Alan Mislove. Picocenter: Supporting long-lived, mostly-idle applications in cloud environments. In 2016 European Conference on Computer Systems (EuroSys '16), London, UK, Apr. 2016.

[8] IBM. Softlayer {Privacy Agreement. https://www.softlayer.com/sites/default/ les/ softlayer privacy agreement - may 2016.pdf. online.

[9] A Deep Learning based approach to VM behavior identification in cloud systems Matteo Stefanini, Riccardo Lancellotti, Lorenzo Baraldi, Simone Calderara Department of Engineering "Enzo Ferrari", University of Modena and Reggio Emilia, Modena, Italy arXiv:1903.01930v1 [cs.LG] 5 Mar 2019.

[10] Virtualization Technology using Virtual Machines for Cloud Computing T. Kamalakar Raju1, A. Lavanya2, Dr. M. Rajanikanth2 1, 2 Lecturer, Dept. of Computer Science, Andhra Loyola College, Vijayawada 3Lecturer, Dept. of Computer Science, Govt. Degree College, Movva. Vol. 4 | Iss. 3 | Mar. 2014.

[11] I. K. Kim, S. Zeng , C. Young,,J. Hwang and M. Humphrey, "iCSI: A Cloud Garbage VM Collector for Addressing Inactive VMs with Machine Learning," *2017 IEEE International Conference on Cloud Engineering (IC2E)*, Vancouver, BC, 2017, pp. 17-28. doi: 10.1109/IC2E.2017.28