# A Novel Security Analysis for Virtualized Infrastructure using Fuzzy Classification Approach in Cloud Computing

R.Siddthan, Dr.A.Nagarajan

Department of Computer Application
Alagappa University, Karaikudi, Tamil Nadu, India.
Sithan314@gmail.com, nagarajana@alagappauniversity.ac.in

***Abstract*— Virtualized infrastructure becomes an attractive goal for cyber attackers for launching advanced attacks in cloud computing. Several existing techniques are utilized for predicting the attacks in the cloud data. It helps to predict the attack effectively and efficiently. But it is difficult to classify the cloud data as normal and attacker's data. Hence a novel security analysis of big data using classification approach is proposed in this work for detecting and classifying the advanced attacks in virtualized infrastructures. Here the logs of the network and user's applications are gathered from the guest virtual machines (VMs). These data are preserved in the Hadoop Distributed File System (HDFS). The process of extracting the features of the attacker is done by using a graph-based event correlation and the possible attack paths are identified based on the Map Reduce parser. After that, the presence of attack can be determined by performing two phase machine learning such as logistic regression and belief propagation. Here the logistic regression can be implemented for calculating the conditional probabilities of an attack regarding the attributes, and belief propagation for calculating the belief in the attack's presence depending upon them. Finally, a fuzzy classification approach is utilized for classifying the normal and attacker's data. The performance of the proposed approach is assessed by utilizing a well-known malware and compared with the prevailing security approaches for virtualized infrastructure. The experimental analysis reveals that our approach performs better in identifying and classifying the attacks with high efficiency and reduced performance overhead**

**Keyword -** Cloud Computing, Big Data, Hadoop and MapReduce, Virtual Machine Security, Fuzzy Classification.

## I. INTRODUCTION

Due to the rapid growth and popularity of processing and storage technologies and also with the success of internet, the computing resources avail more ubiquitously and cheaper. It is then referred to as cloud computing which offers the requirements of present and upcoming information and communication technology [1]. Cloud computing is recognized as a model that offers computing resources depending upon the pay-per-use by configuring such resources dynamically for accommodating several needs of workload. This can be done by the exploitation of virtualization [2]. Virtualized infrastructure is comprised of virtual machines (VMs) which depends upon the software-defined multi-instance resources of the hosting hardware. The software-defined multi-instance architecture is managed, sustained and regulated by the virtual machine monitor which is also referred to as hypervisor. The extensive utilization of virtualized infrastructures becomes a substantial provision for cloud computing services due to the facility of pooling various computing resources in addition to empower the on-demand resource scaling [3]. This makes the virtualized infrastructures as an interesting goal for cyber attackers to get illegal access by launching attacks. Some degrading attacks such as Virtualized Environment Neglected Operations Manipulation (VENOM) is accomplished for the exploitation of software issues in the source code of hypervisor. It helps an attacker to escape from a guest VM and access the essential hypervisor. Also the issues in the operating system due to the attacks like Heartbleed and Shellshock can be utilized against the virtualization infrastructure for the purpose of obtaining the login informations about the guest VMs and performing attacks to Distributed Denial of Service (DDoS).

DOI: 10.21817/ijet/2018/v10i4/181004043

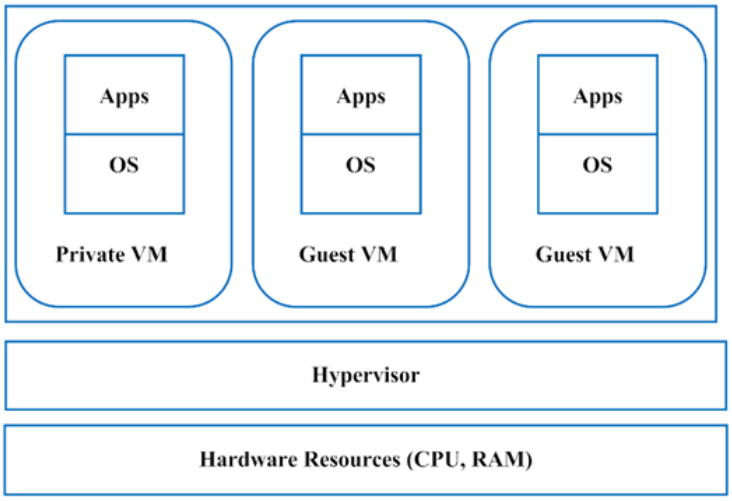

Fig. 1 Virtual Machine Security in cloud computing

The attack traffics are of two main categories such as IP spoofing attacks and real source IP-based attacks. Generally the real source IP based DDoS attacks uses the negotiated nodes called as bots or zombies for the purpose of launching an attack. Conversely, IP spoofing DDoS attacks are utilized on energy optimization. In existing systems Distributed Denial of Service (DDoS) attacks are continued as the highest security concerns due to the constant increase on the volume of DDoS attacks. Among them, the most common type of attack is the SYN Flood attack. Because of the increased capability of hardware resources, the traditional DDoS security solutions are not superior which results in increased cost and long utilization cycle. Hence the new opportunities are introduced for reducing the amount of proprietary hardware which is required for launching and operating the network services by an emerging cloud Network Function Virtualization (NFV) [4].

Here a novel security analysis for big data is proposed in this work for protecting the virtualized infrastructure from the advanced attacks. The main contribution of this research work is to identify and classify the data. For this purpose a fuzzy classification approach is utilized in this work. The results reveals that our suggested solution can classifies the data as normal or attacker’s data in an effective manner. The remaining parts of this paper is systematized as follows: Section II debates the relevant topics of this work. Our proposed framework is discussed in Section III. The evaluation results of the proposed methodology is presented in Section IV. Finally, concluded the effectiveness of this proposed work in Section V.

## II. Related Work

[5] reviewed the rise of big data and their issues in cloud computing. In this study, a model was proposed for the classification, conceptual view and cloud services of big data and also it was compared with the various existing representative platforms of big data. Also the background and the core elements of Hadoop technology like MapReduce and HDFS. The challenges in the process of big data in the measures like scalability, data integrity, volume, availability, data access, generation, protection and transformation. Also the issues related to the privacy and legal problems were studied. The main drawback in cloud computing was data management. [6] introduced a first protocol namely SecCloud which helped to achieve privacy cheating hindrance by utilizing the designed verifier signature, batch verification and probabilistic sampling approaches. The main contribution of this research work was to reduce the cost by obtaining an optimal sampling size. Also the implementation of SecCloud was carried out by building a practical secure aware cloud computing which was also referred to as SecHDFS. The results showed the superiority and efficiency of the proposed SecCloud. But the major drawback was their detailed computations such as data mining and linear programming computation. And also they had some issues related to the privacy preserving.

[7] focused on the threat like co-resident attack in which the malicious users created the side channels and mined the private data from the virtual machine that were co-located on the same server. The main motive of this work was to eliminate the side channels. Hence to solve this issue, a policy for improving the virtual machine allocation was studied. It made the attackers hard to co-locate their targets. Also a novel policy was implemented to balance the workload and to reduce the power consumption. [8] discussed about the significant features of the cloud and their service delivery and deployment models. Also its causes for adoption and the problems that obstruct their broad adoption. The three cloud security frameworks like ENISA, CSA and NIST were surveyed, offered a review about the security, trust issues, privacy and helped to understand the present status of the cloud security. Also a wide range of security and privacy concerns were presented for identifying the major threats and classified them to six groups. This helped to address the issues and also it was essential to remove these issues.

[9] discussed about the issues and the possible solutions to defend the big data in cloud computing. Then a Meta cloud data storage was proposed for guarding the big data in the environment of cloud computing. Here the number of users in the cloud data center was found by utilizing the Map Reduce framework. The protection of mapping several data components for each provider was carried out using met cloud data storage interface. The main benefit of this proposed approach was that even though it required high execution effort, it offered valuable data for cloud computing which would have an effect on the next generation systems. [10] deliberated the security concerns in big data, cloud computing and Hadoop environment. The main intention of this research was focusing the security issues associated with the big data in cloud computing. Also several possible solutions for the problems in Hadoop and cloud computing security were discussed. Moreover the applications, benefits and drawbacks of big data and cloud computing were represented in this work. It would be helpful for the upcoming frontiers in science and technology. Also the proposed approaches were utilized for security in complex business operations in the cloud computing.

[11] referred the Scientific Data Infrastructure (SDI) model and the security service linked with the anticipated Federated Access and Delivery Infrastructure (FADI) were discussed. Also this work offered recommendations for the practical realization of significant elements of security infrastructure such as fine grained data centric access control policies, federated access control and identity management and Dynamic Infrastructure Trust Bootstrap Protocol which helped to deploy the situation of remote virtualized data processing which were trusted. It would be helpful for the process and infrastructure of bigdata by the progress of respective Cloud and InterCloud architecture framework. [12] introduced a large iterative multitier ensemble (LIME) classifier for analyzing the security in big data. This classifier was large but they were easy to use and generate. This classifier combined various ensemble Meta classifiers into many tiers simultaneously and integrated then into one system which generated iterative system automatically. Here the performance of the LIME classifiers were investigated for the issues regarding security of big data. The results showed that the LIME classifier increased the classification accuracy significantly and also it performed well compared to the base classifiers and standard ensemble Meta classifiers.

[13] presented a system which utilized the virtualization technology for the energetic assignment of data center resources depending up on the difficulties of application and support of green computing with the optimization of the number of servers that were utilized. Here the concept of skewness was introduced for measuring the unevenness in the utilization of multi-dimensional resource of a server. Various types of workloads could be combined and the overall utilization of server resources could be improved by reducing the skewness. Also the overload in the system could be prevented effectively with the development of heuristics in the utilization of saving energy. [14] explored the performance of the conventional virtual machine and the Linux containers. For comparing the performance of both VM and Linux, the workloads that stress the memory, utility, CPU, storage and resources for networking was utilized. The results depicted that both VM and Linux were need to tune the provision of I/O intensive applications. The main drawback was that in this work, single VMs or Linux were created for processing the entire server. This leads to more investigations in the measures like performance isolation for multiple workloads on same server, tradeoffs amongst the scale-up and scale-out and restarting and live migration.

[15] focused on the co-resident attack in which the malicious users were aimed to co-locate the VMs with the targeted VMS on the same server. Then the private data were extracted by exploiting the side channels from the victim. In this work, the problem was viewed from the different perspective and investigated to reduce the probability of co-locating VMs with the targets with the maintenance of acceptable workload and power consumption for the system. Different allocation policies of VM were compared by introducing a security game model. The results demonstrated that the cloud provider reduced the probability of attaining co-location by the deployment of a policy pool instead of a single policy in which the individual policies were selected depending up on certain probability. But this system was suitable for large scale systems and also it hard to identify the behavior of attacker and normal users in this mixed policy. [16] identified the technical issues in the support of real time applications in cloud and also surveyed the current advancements in practical virtualization and cloud computing technology. The presented concerns range from the structure and role of hypervisor, different possible virtualization types and their corresponding performance, resource management and scheduling of VMs with the hierarchical scheduling and the significant part of the virtualization technology. The bottlenecks in the execution of protocol in various layers of the software stack were removed by describing some solutions of HPCC community. To connect the cloud and distributed real time systems, a terminology was mapped between them.

[17] explored the issues related to the security in virtual infrastructures for determining the existing threats and complexities in the cloud platform. The infrastructure and the problems in cloud virtual systems were discussed with the traditional security approaches. Finally some key challenges in the research for implementing the new virtualization aware security solutions were proposed and explored. It offered pre-emptive protection for ever-dynamic and complex virtual infrastructure in cloud. These solutions had the ability

to offer detection and protection of both known and unknown threats in real time applications. [18] overviewed the big data technologies and cloud computing. Specifically the technologies of big data like Apache Hadoop was also discussed in this survey which offered the parallelized and distributed data processing and the analysis of petabyte scale data together with the review of current Hadoop utility in the community of bioinformatics. The disadvantage of this was that most of the cloud infrastructures offered small ability on the application, data and service interoperability which caused the customer hard to migrate from one provider to other provider. [19] introduced five different virtualized cloud computing servers (VCCS) and provided the suitable evaluation for five renowned hypervisors in VCCS. Here the analysis of the performance of virtualized server and shared storage accessing along with the calculation of consolidation ratio and TCO/ROI in the utilization of server were taken. The performance of VM achieved equal level for all hypervisors but the calculation of VM TCO/ROI and density were different between them, when a scheme was essential with better ROI and least TCO, the ESX server was recommended in the utilization of server at last.

## III. Proposed Method

This section demonstrates the implementation of security analysis for big data for detecting and classifying the advanced attacks in virtualized infrastructures. Network logs in addition to the application logs of the user are gathered periodically from the guest virtual machines (VMs) and kept in the Hadoop Distributed File System (HDFS). The large volume of DDoS attacks can be handled by CoFence which permits the domain networks to help through resource sharing. The flow of the proposed methodology is shown in Fig. 1.

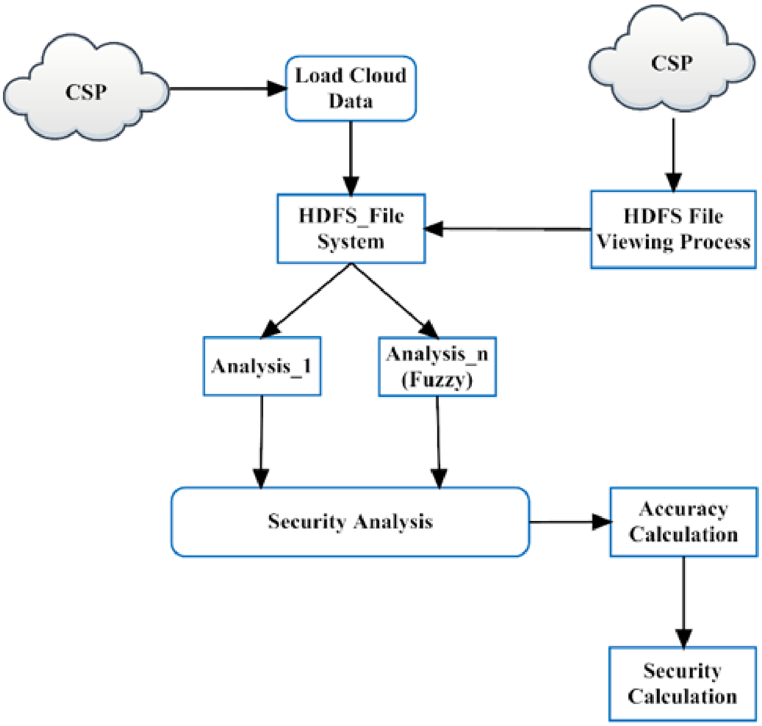

Fig. 1 Work flow of the proposed approach

Initially the cloud service provider loads the cloud data. Then the cloud service provider creates a Virtual Machine (VM). The data are collected and kept in Hadoop Distributed File System (HDFS). After this the analysis of data is carried to find which data are normal and malicious. After this Fuzzy classification approach is utilized for categorizing the data as normal and attacked. Finally the classification and security is evaluated.

### *A. Cloud Service provider Login and Load data*

The cloud service provider can handle all the processes in the system. An individual can get the access to a system by logging in or signing in for the purpose of authentication. The process of login is carried out in the form of “username” and a matching “password”. The recent secure systems frequently necessitate a second aspect for extra security. The cloud data can be loaded by the cloud service provider which may consists of several attributes such as name of the customer, cloud cost of the customer id, name of the server and VMs. The data that are stored in the cloud are huge and hence the concept of Hadoop-Distributed-File System (HDFS) is used in this work. This helps to secure the stored data for further analysis of data. The loaded data is then stored into a cloud server and transfers the large amount of data into big data storage.

### B. Host and VM Creation

After loading the data, the cloud server creates a Virtual Machine (VM). Then the famous taxonomy of scheduling algorithm is utilized in the distributed computing systems for obtaining the quality of solutions. The scheduling algorithm is classified as two types such as optimal or sub-optimal. In optimal scheduling approach, the data can be characterized depending upon the entire information about the distributed environment such as abilities and load of the hardware along with the requirements of the resources. Whereas, when the information is not available or when the system has infeasible time for computation, the Sub-optimal scheduling is used.

### C. Data Storage and Analysis

The large amount of data can be stored in HDFS which helps to secure the data for further process. Here, the MapReduce is utilized for storing and securing the data. It is a programming model for producing big data sets. A Map Reduce program consists of a Map() procedure which accomplishes the operation of organizing and categorizing and a Reduce() method which accomplishes a summary procedure. Then the stored data are analyzed for determining whether the loaded data are secure or attacked. The data analysis is the process of examining, altering, and exhibiting for determining some beneficial information, recommending inferences, and associating decision-making. In this approach, the process of data analysis is mainly focused on modeling and knowledge discovery for predictive measures.

### D. Malware Prevention

From the stored and processed data, the malware detection and is carried out. The data can be analyzed for determining the normal data and malicious data. After that the classification approach is utilized for categorizing the normal and attacked data. Generally, the analysis does not depends upon a-priori attack signatures and also it does not deliberate information about the payload, but instead of that it depends upon the per-flow meta-statistics which is determined from packet header and volumetric information such as counts of packets, bytes, etc. However, our scheme has the ability to function with signature-based methodologies on an online basis in the circumstances where the decryption is achievable and cost effective. The main intention of this proposed approach is to implement the detection techniques that are precisely directed towards the cloud and incorporate with the infrastructure for not only detecting but also offering flexibility through remediation. Thus the fuzzy classification is utilized in this work for identifying the attackers and normal event.

## IV. Performance Analysis

This section demonstrates the performance of the proposed methodology by using various performance measures such as accuracy, precision and efficiency. Also our proposed methodology evaluates its performance and compared with the existing techniques.

Accuracy

Accuracy can be evaluated as the correctness and efficiency of the classification techniques. It determines how accurately the data is classified using classification approach. It is evaluated based on the formula,

Accuracy= (TP+TN)/((TP+TN+FP+FN) )

Precision

The most widely utilized measure for determining the performance of data classification is precision. It offers relevant results in data classification and it is given by,

Precision= TP/((TP+FP) )

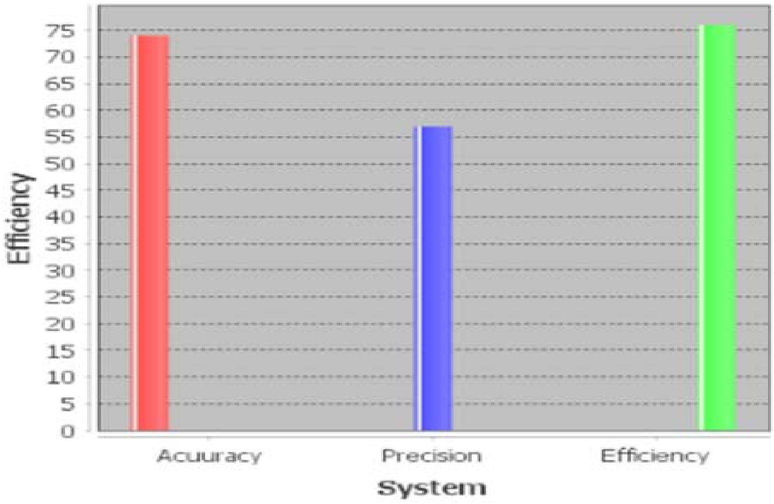

Fig. 3 Performance Analysis

Fig. 3 depicts the performance analysis of the proposed fuzzy classification approach. It measures the accuracy, precision and efficiency of the system. The suggested technique provides improved accuracy results for data classification. From the results it is observed that the anticipated approach offers superior results with higher efficiency.

Fig. 4 illustrates the comparative analysis of the efficiency of both the proposed and existing approaches. From the results it is noted that the proposed approach offer higher efficiency in classification of data as normal or attacker when compared to the existing approach.

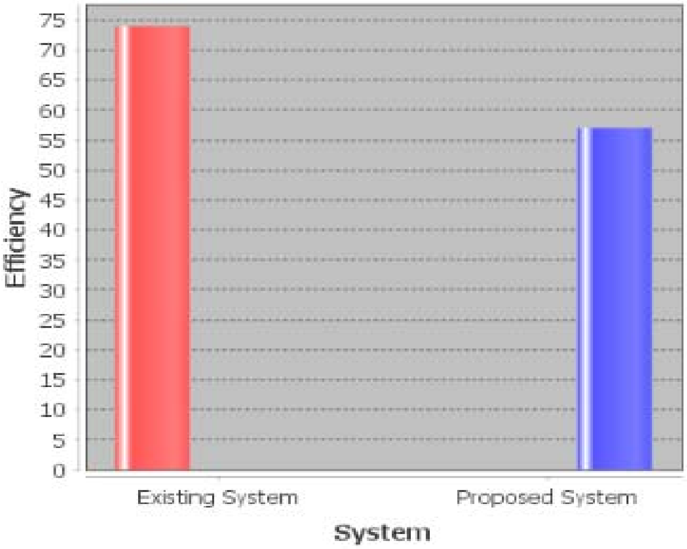

Fig. 4 Comparative analysis

The utilization of Virtual machines for both the existing and proposed approaches are evaluated and shown in Fig. 5. The proposed approach uses less number of virtual machine compared to that of the existing approach. This proves the superiority of our proposed approach.

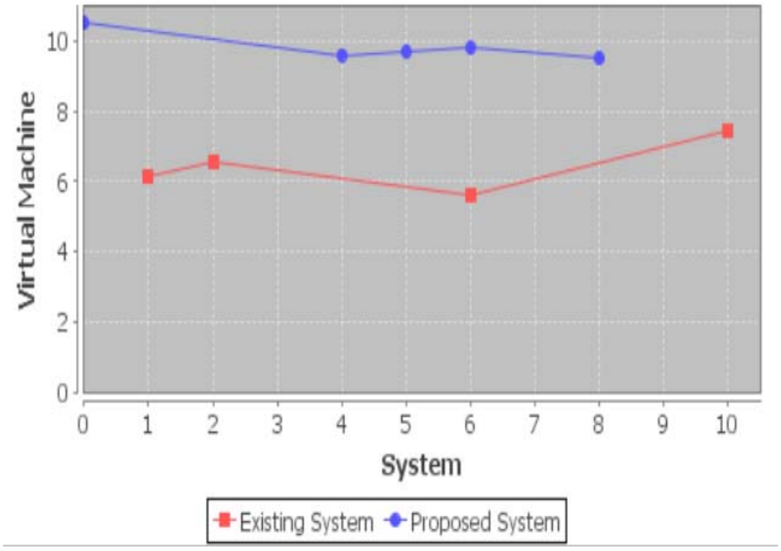

Fig. 5 Utilization of Virtual Machine

Fig. 6 represents the execution time taken for both the proposed and existing approach. The graph shows that the performance of the proposed approach takes less time than the existing system. From this, it is analyzed that the proposed approach performs better than the existing technique.

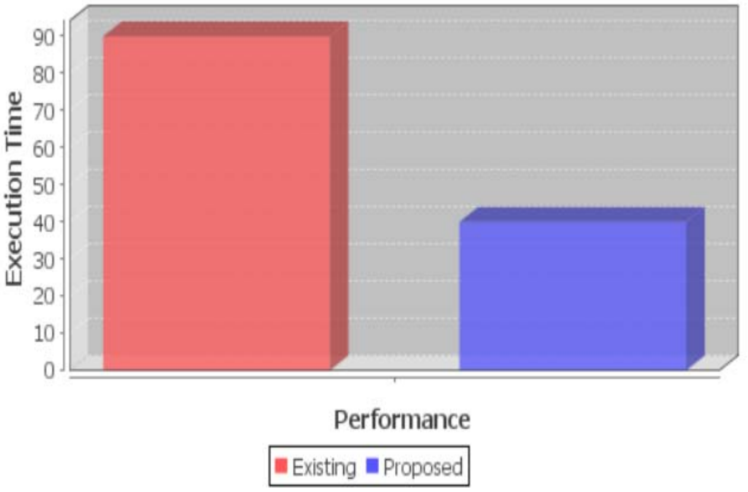

Fig. 6 Execution time

## V. CONCLUSION

In this work, a novel security analysis for big data is proposed for securing the virtualized infrastructures in cloud computing in contradiction of advanced attacks. Initially, the network logs of the guest VMs in addition to the application logs of user are gathered from the guest VMs and kept in the HDFS. Then, the features of the attacked data are extracted by correlation graph and Map Reduce parser. Finally, two phase machine learning is exploited to determine the presence of attack. Logistic regression is realized for determining the conditional probabilities of attack regarding the attributes of an individual. Finally the fuzzy classification approach is used for identifying the normal and attacker’s data. From the experimental analysis, it is concluded that the suggested method gives better data classification results when compared to the existing approaches.

## REFERENCES

[1] D. Puthal, B. Sahoo, S. Mishra, and S. Swain, "Cloud computing features, issues, and challenges: a big picture," in Computational Intelligence and Networks (CINE), 2015 International Conference on, 2015, pp. 116-123.
[2] I. Pietri and R. Sakellariou, "Mapping virtual machines onto physical machines in cloud computing: A survey," ACM Computing Surveys (CSUR), vol. 49, p. 49, 2016.
[3] T. Y. Win, H. Tianfield, and Q. Mair, "Big data based security analytics for protecting virtualized infrastructures in cloud computing," IEEE Transactions on Big Data, vol. 4, pp. 11-25, 2018.
[4] W. Rankothge, J. Ma, F. Le, A. Russo, and J. Lobo, "Towards making network function virtualization a cloud computing service," in IM, 2015, pp. 89-97.
[5] I. A. T. Hashem, I. Yaqoob, N. B. Anuar, S. Mokhtar, A. Gani, and S. U. Khan, "The rise of “big data” on cloud computing: Review and open research issues," Information Systems, vol. 47, pp. 98-115, 2015.
[6] L. Wei, H. Zhu, Z. Cao, X. Dong, W. Jia, Y. Chen, et al., "Security and privacy for storage and computation in cloud computing," Information Sciences, vol. 258, pp. 371-386, 2014.
[7] Y. Han, J. Chan, T. Alpcan, and C. Leckie, "Using virtual machine allocation policies to defend against co-resident attacks in cloud computing," IEEE Transactions on Dependable and Secure Computing, vol. 14, pp. 95-108, 2017.
[8] T. Islam, D. Manivannan, and S. Zeadally, "A classification and characterization of security threats in cloud computing," Int. J. Next-Gener. Comput, vol. 7, 2016.
[9] G. Manogaran, C. Thota, and M. V. Kumar, "MetaCloudDataStorage architecture for big data security in cloud computing," Procedia Computer Science, vol. 87, pp. 128-133, 2016.
[10] V. N. Inukollu, S. Arsi, and S. R. Ravuri, "Security issues associated with big data in cloud computing," International Journal of Network Security & Its Applications, vol. 6, p. 45, 2014.
[11] Y. Demchenko, C. Ngo, C. de Laat, P. Membrey, and D. Gordijenko, "Big security for big data: Addressing security challenges for the big data infrastructure," in Workshop on Secure Data Management, 2013, pp. 76-94.
[12] J. H. Abawajy, A. Kelarev, and M. Chowdhury, "Large iterative multitier ensemble classifiers for security of big data," IEEE Transactions on Emerging Topics in Computing, vol. 2, pp. 352-363, 2014.
[13] Z. Xiao, W. Song, and Q. Chen, "Dynamic resource allocation using virtual machines for cloud computing environment," IEEE Trans. Parallel Distrib. Syst., vol. 24, pp. 1107-1117, 2013.
[14] W. Felter, A. Ferreira, R. Rajamony, and J. Rubio, "An updated performance comparison of virtual machines and linux containers," in Performance Analysis of Systems and Software (ISPASS), 2015 IEEE International Symposium On, 2015, pp. 171-172.
[15] Y. Han, T. Alpcan, J. Chan, and C. Leckie, "Security games for virtual machine allocation in cloud computing," in International Conference on Decision and Game Theory for Security, 2013, pp. 99-118.
[16] M. García-Valls, T. Cucinotta, and C. Lu, "Challenges in real-time virtualization and predictable cloud computing," Journal of Systems Architecture, vol. 60, pp. 726-740, 2014.
[17] A. S. Ibrahim, J. Hamlyn-Harris, and J. Grundy, "Emerging security challenges of cloud virtual infrastructure," arXiv preprint arXiv:1612.09059, 2016.
[18] A. O’Driscoll, J. Daugelaite, and R. D. Sleator, "‘Big data’, Hadoop and cloud computing in genomics," Journal of biomedical informatics, vol. 46, pp. 774-781, 2013.
[19] B. R. Chang, H.-F. Tsai, and C.-M. Chen, "Evaluation of virtual machine performance and virtualized consolidation ratio in cloud computing system," Journal of Information Hiding and Multimedia Signal Processing, vol. 4, pp. 192-200, 2013.