Data Analytics and Artificial Intelligence
Vol: 1(3), 2021
REST Publisher
ISBN: 978-81-948459-4-2
Website: http://restpublisher.com/book-series/daai/

Securing the Future: Enhancing Cloud Computing Security and Data Privacy for Big Data and Virtualization Infrastructure

Anusha Prem I
St. Joseph's College of Arts and Science for Women, Hosur, Tamil Nadu, India.
Corresponding Author Email: ianushaprem@gmail.com

Abstract

Data-centric resources will eventually occupy a larger portion of the technology landscape. The cloud computing and big data infrastructures will need to be more secure. Recent technological advancements have shown how important data is to nearly controlling every facet of an organization's operations. Cloud computing has considerably improved the data, its privacy, and the execution of many Big Data applications. This essay discusses cloud computing's applications, security, and privacy in building a secure framework for big data architecture and virtualization. Security protocols for several linked Data Science domains is an exciting field of computer science research. Keywords—Big Data, Cloud Computing, Virtualization

1. Introduction

Cloud computing, big data, and virtualization are the three data-driven platforms that have developed since the Internet's inception. They still hold a monopoly on the management, exchange, and preservation of data for numerous large-scale businesses as well as smaller ones. The terms "cloud computing" and "the cloud" refer to the availability, scalability, agility, and cooperation. It allows for cost savings achieved through effective and efficient computing [1]. The National Institute of Standards and Technology (NIST) has categorized the Cloud Computing environment into three services models, four deployment methods, and its basic characteristics [2]. An enterprise's reliance on the Cloud for its operations increases its susceptibility to security breaches and assaults. Big Data, as the name suggests is large amounts of data collected, processed and stored. The Big Data is classified based upon the four V's abbreviation. The four V's are: Volume, Velocity, Variety and Veracity [3]. Big Data analytics contain metadata and can be used to expose the privacy of an individual or any organization, security measures for the same need to be constructed. Infrastructure for virtualization and the related technologies are becoming quickly recognized in the market. The process of creating a virtual computer, or virtualization, results in the construction of two operating systems on one operating system. One Linux distribution that can be loaded on a virtual machine platform such as VMware [5] or Oracle Virtual Box [6] is Fedora [4]. This allows the Linux-based operating system to be used on a computer that is running Microsoft Windows. This paper's main goal is to raise awareness of the security measures that are currently in place for these information processing platforms and to propose some ideas that can be put into practice to further secure and protect the data while preserving its consistency and integrity.

SECURITY OF THE CLOUD

"Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction," according to the National Institute of Standards and Technology (NIST) [7].*Cloud Platforms* The Cloud Computing platform can be further categorized by the services it offers and the models on which it can be deployed [8], [9]

- *Infrastructure as a Service (IaaS)*

This platform gives customers access to network connectivity, storage, and virtualized resources. These resources can be scaled by users as needed. One possible usage for IaaS is as a high-end cloud system. Typical instances of these services are Microsoft Azure, GoGrid, and Amazon EC2.

- *Platform as a Service (PaaS)*

It is the platform that offers virtualized resources for development with improved aspects of the programming platform. Additionally, it facilitates the geographical spread of developers who work on the platform's development. Among the PaaS examples are Heroku, Google App Engine, and Amazon Map Reduce/Simple storage.

- *Software as a Service (SaaS)*

The most sophisticated cloud computing platform, it offers on-demand access to services that are mostly found on computer networks or web browsers. In addition to being more regularly available, there is no need to buy a license. The SaaS platform is readily available, making it simple to combine with mashup applications. Some well-known examples are Salesforce.com and Google Maps.

- *Proposed Security Architecture*

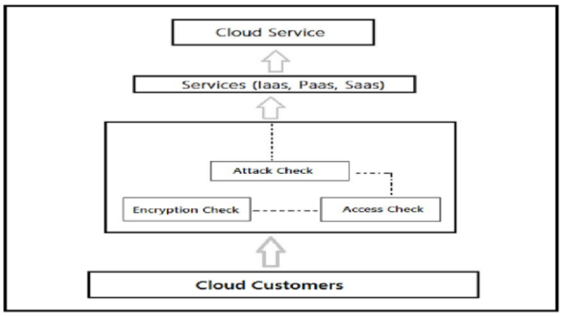

FIGURE 1. Proposed security architecture:- Cloud

The proposed architecture for securing the Cloud Computing infrastructure contains four levels and several components (Fig 1). The distribution of the various components over the architecture promise to provide multiple levels of security and privacy once in implementation.

Layer I:

Customers and businesses who use the cloud as their storage medium are supported by this tier. This covers database accessing tools, graphical user interfaces, and front end devices. The end user or company can connect to a network and, ultimately, their own Cloud storage through the apps provided in this layer.

Layer II:

This layer contains the main security mechanism of the architecture. There are three levels of filterations as follows:

- *Encryption Check*

The user is required to input a unique key that is produced at random and sent to them when they access the Cloud via one of the accessible platforms (IaaS, PaaS, or SaaS). This key serves as a tool for private key encryption. Users are forwarded to the second filtering layer if the key matches. It is impossible to go past this layer unless the attacker has remote access to the randomly generated key.

- *Access Check, Data access control*

The second layer of filtration is this one. The attackers are unable to go above this layer, even if they are able to obtain full access to the unique key. The users will only be able to access the data that they need, rather than all of the data, depending on the service they have chosen. This aids in preserving the data's secrecy and integrity. Regarding national security, the cloud storage may contain a number of data that must be classified for a limited number of people to view. The details of every single account holder for a cloud-based financial service (account number, amount). Users should be directed to the appropriate Cloud storage space based on their kind in order to make the necessary data available. To improve the digital signature check, users should be required to authenticate themselves using facial recognition software or by keeping track of their typing pace. They are then given access to the Cloud services after that.

- *Vulnerability/Attack check*

The final stage before allowing access to the chosen service is to determine whether the cloud is susceptible to any attacks, having previously verified the user's identity and access privileges. At this stage, an attack could expose the integrity and confidentiality of the user's profile and data. To ensure the highest level of protection for the cloud environments used by the government, private companies, banking, healthcare, and, lastly, individual accounts, the utmost care must be taken.

Layer III:

It maps the chosen service to the user's supplied data and grants them unrestricted access to it by navigating past several security checks and the second layer. The system administrator keeps track of and logs the timestamps associated with the allocation of Cloud access and the time of departure. If these records are kept for forensic analysis, they will be useful. The suggested architecture may appear a little complicated to construct, but it will offer numerous layers of protection, assisting in the preservation of the data's consistency, integrity, and privacy. The confidentiality of the users' or businesses' accounts will undoubtedly be protected by adding and carrying out these security measures at every stage, from connecting to the network to using the cloud.

THE BIG DATA FRAMEWORK

The Ministry of Defense provides the following definition of big data technology: "Big Data - large pools of data that can be captured, communicated, aggregated, stored, and analyzed - is now part of every sector and function of the global economy." This definition helps to clarify the true nature of big data. It is becoming more and more clear that data is necessary for a large portion of contemporary economic activity, innovation, and growth—much like other critical components of production like tangible assets and human capital” [10].

UNDERSTANDING BIG DATA

Big Data analytics use four distinct characteristics to define the process and magnitude in which the data is generated and processed. The generic abbreviations include the four V’s.

- Volume:

The mass quantities of data that is being generated on a daily basis.

- Variety:

The variety within the data generated, is all about manag- ing the complexity of multiple data types. This includes both structured and unstructured data.

- Velocity:

How quickly information is moved or distributed from point A to point B. This value is used to calculate how long it will take to refresh. Due to the dynamic nature of the data, additional processing power, precise design, and excellent, dependable, and quick network speeds are needed.Veracity:

The amount of reliability with respect to the data is the veracity. The current view in the industry is of the precision of individual data items. And, the loss of relevance caused due to the normalizing effect of analyzing the data, in terms of its huge magnitude.

SECURING THE BIG DATA INFRASTRUCTURE

While considering the security measures for the Big Data infrastructure, it is necessary to classify the existing architecture. Big Data is different in terms of its architecture and its operation. The architectural sector supports the data and the database security. The operational sector handles the challenges in securing big data and highlights the deficiency in these systems.

Architecture:

Big Data, is classified by its deployment model, consisting of highly distributed, redundant, elastic data repositories supported by the Hadoop file system. [11]

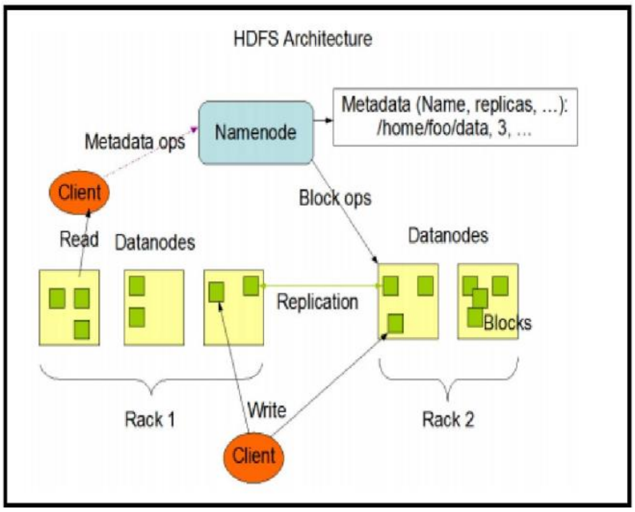

FIGURE 2. HDFS Architecture

It is easiest to comprehend the Big Data architecture by using the Hadoop Distributed File System (Fig 2). It is software that enables the distributed processing of big data sets across computer clusters and is powered by the Apache foundation. It offers application data access at a high throughput. Numerous factors, such as dispersed nodes, sharded data, data ownership, inter-node communication, etc., are included in architectural security. The Big Data architecture is connected to the security precautions described in [11]. Big Data's distributed file system and high volume of data are the main causes of security and privacy issues. The architecture's capabilities must correspond with the data produced.

Operation:

While considering the security and privacy measures for the Big Data interface, we need to consider securing both the data and the infrastructure. The various factors which should be considered while analyzing operational security are:

- Stored Data

The data which is being stored onto the Cloud is most likely to be infiltrated with. This data can be best protected using encryption. Encryption helps to protect the data which is copied from the cluster.

- Administrative Access

Any enterprise which intends to backup their critical data on Big Data servers, needs to hand over the encryption mechanism into the hands of a trustworthy system admin- istrator. Unwanted access to data can be alarming, making data more vulnerable.

- Auditing and maintaining Logs

If there is probable breach into the cluster, then the enterprise should be able to realize and analyze it. The most effective way of auditing is by maintaining logs, each time the data is accessed, read, retrieved or modified. This can be achieved by integrating the database with open source logging tools.

As we have seen the various parameters on which the Big Data infrastructure is established and expanded, one thing can be confronted is that, failures arise from complexity. Cryptography, auditing, and other mechanisms may prove to be few of the solutions to protect the data, but in reality, combining the data generated and pressures from push web applications, it should be clear that problems are due to arise.

SECURITY OF VIRTUALIZATION

The idea behind virtualization is to establish a workstation that can support the operation of two separate operating systems. A virtual machine (VM) can be used to implement virtualization. A hypervisor known as the Virtual Machine Monitor (VMM) is used to further monitor the virtual machines that are built [12]. A guest operating system is an operating system that runs within a virtual computer. Several operating systems can run simultaneously on a single host computer thanks to hypervisors. the hypervisors set up on the hardware of a server. Only the operating systems of the visitors are used.

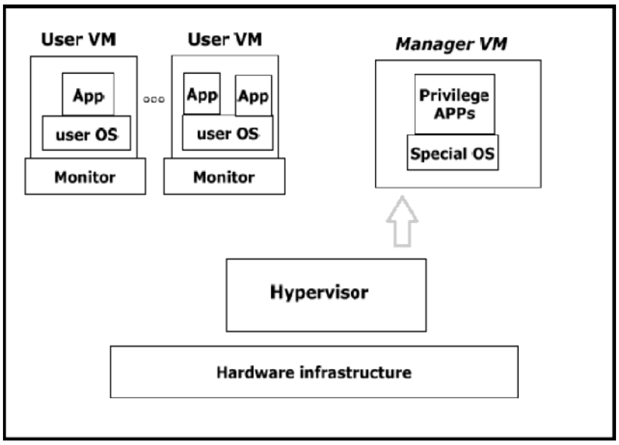

FIGURE 3 Existing security architecture

Hypervisor based Virtualization, (Fig 3) functions using defense mechanisms like intrusion detection systems, or con- trols the virtual machines through the partitions created during booting. Some of the VMs contain privileged partitioning which view and control the VMs. This model is partly se- cure as it puts the hypervisor at the center of the privacy mechanism. If infiltrated upon, the hypervisor makes all of the virtual machines running vulnerable to attacks and a breach in security [13]. Considering this vulnerability a more secure architecture is proposed.

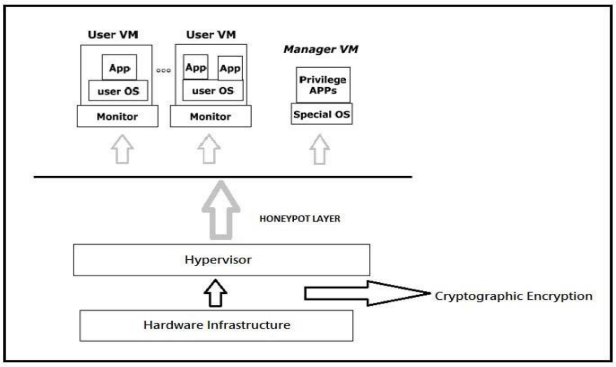

FIGURE 4. Proposed security architecture:-Virtualization

The architecture which we have proposed uses a two layered security mechanism. (Fig 4)

Layer I

Between the hardware of the machine on which the virtual machines are running and the hypervisor, a cryptographic encryption layer is introduced. The entire hardware is encrypted so as to protect the contents within the hard drive. This layer protects the contents of the hardware from malicious attacks even if the intruder is able to bypass the hypervisor. The data within the hard drive is secure as it cannot be accessed without the decryption key.

Layer II

In between the hypervisor and the virtual machines a Honeypot layer is deployed [14]. Honeypots are used to create a fake en- vironment in which the attackers feel that they are penetrating into the real servers but in reality their modus operandi is being recorded. Thus, even if the attackers penetrate through the initial layer, they fall into a trap in the second layer. The VMs still possess a threat of being vulnerable, but in this case, the details of the attacker can be recorded and retraced. Thus adding a secure measure to the infrastructure.

These two levels of protection will provide a stable security and privacy to the virtualization infrastructure.

2. CONCLUSION

It has been determined that these three commonly utilized data/information processing platforms represent the future of computer science after taking into account the various security and privacy paradigms. Petabytes of data are produced every day by the modern world and the Internet, which controls the majority of activities through the World Wide Web.Future research areas that will be most concentrated will be those that work in the same fields. More data will be produced as science and technology develop and more people utilize the internet, which will lead to an increase in user numbers. More individuals are becoming familiar with this intriguing technology as a result of the movement in the digital world toward free and open-source operating systems and the Internet. Future research efforts will be primarily focused on governing this data and creating a secure foundation for it. In this paper, we have attempted to discuss the security mechanisms that are already in place and suggest a few approaches to improve upon them. Thus, we may safeguard data, the most important component of the digital revolution, by enhancing security measures. safeguarding the infrastructure of virtualization, big data, and the cloud in the end.

References

1. C. Security, “https://cloudsecurityalliance.org/csaguide.pdf.”
2. NIST, “http://www.nist.gov/itl/cloud/upload/cloud-def- v15.pdf.”
3. R. A. E. Ahmed, “A survey of big data cloud computing security,” in IJCSSEInternational Journal of Computer Science and Software Engineering, 2014.
4. Fedora, “https://fedoraproject.org/wiki/overview.”
5. VMWare, “https://en.wikipedia.org/wiki/vmware.”
6. Virtualbox, “https://www.virtualbox.org/.”
7. N.I. of Standards and Technology, “csrc.nist.gov/publications/nistpubs/800-145/sp800-145.pdf.”
8. W.Tsai, “service-oriented cloud computing architecture,” in Seventh International Conference on Information Technology, 2010.
9. Securosis, “https://securosis.com/assets/library/reports/securingbigdata.”
10. Rusi,“https://www.rusi.org/downloads/assets/rusi bigdata report 2013.pdf.”
11. Hadoop, “https://hadoop.apache.org.”
12. Hypervisor, “https://en.wikipedia.org/wiki/hypervisor,” in online.
13. F. Sabahi., “Secure virtualization for cloud environment using hypervisor-based technology,” in International Journal of Machine Learning and Computing, 2012.
14. Honeypot, “https://en.wikipedia.org/wiki/honeypot.”