IJARASEM

ISSN: 2395-7852

International Journal of Advanced Research in Arts, Science, Engineering & Management

Volume 10, Issue 5, September 2023

ISSN
INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

Impact Factor: 6.551

☎ +91 9940572462 +91 9940572462 ✉ ijarasem@gmail.com @ www.ijarasem.com

Big Data Security Tools and Management Practices for Cloud Environment

AGGALA CHIRANJEEVI, DR. PRASADU PEDDI, DR. SUNEEL PAPPALA

Research Scholar, Dept. of Computer Science & Engineering, Shri JJT University-Rajasthan, India

Research Supervisor, Dept. of Computer Science & Engineering, Shri JJT University- Rajasthan, India

Co-Supervisor & Associate Professor, Dept. of Computer Science & Engineering, Lords Institute of Engineering and Technology, Hyderabad, India

ABSTRACT: The IT industry has recently grown enamoured with concepts of "cloud computing" (the use of the Internet to store and manage vast amounts of data) and "Big data" (the challenges of dealing with large, unstructured information). This study aspires to demonstrate the efficacy of secure cloud analytics in protecting against prevalent virtualization-based threats such denial-of-service attacks, SQL injections, cross-site scripting attacks, malicious URLs, probes, and user-to-hostname/hostname-to-user lockout attacks. Following the successful introduction of MapReduce-based secure analytics for virtualized infrastructure, this study evaluates the effectiveness of utilising up-to-date AES to partition and encrypt cloud-based cubes. Using Hadoop and R-Programming, this experiment uncovered IDS. The second approach included using the Hadoop Framework to detect and simulate complex attacks in a controlled lab environment. Finally, we utilised Cloud Sim to provide a simulation technique to developers of virtual infrastructure and big data, creating a broad variety of VMs that communicate with one another and exhibit unusual behaviour (in the case of certain VMs, we infected them with malware and viruses). Our simulation shows that cloud-based threat detection boosts performance in a wide variety of scenarios.

KEYWORDS: Big data, Hadoop Framework, SQL injections, cloud computing, MapReduce, Cloud Sim.

INTRODUCTION

Cloud computing (CC) systems may be defined while concealing underlying structure by using computer technology that pools the processing capacity of several interconnected devices. The process of developing the networks that people use today is a prime example of this. While CC has been available for some time, the options it brings for app and service development that allow its utilization in the back office have caused a significant change in how IT personnel utilize it. An environmental issue warrants attention. Using too much energy (electricity) and time (not handling matter) to keep a host environment at a comfortable temperature is inefficient. It's feasible that many companies won't need this in future and will find ways to avoid it. Managers of data centers have significant challenges related to energy efficiency. One of the most difficult tasks facing managers is controlling costs, especially in the latter few years of a person's life when health problems begin to arise. The data centers' massive need for electricity is staggering. This implies that they are located in close proximity to both power plants and sources of electricity production. To keep the machines working, they need a large infrastructure, particularly considering the high levels of isolation and energy consumption. To facilitate the global dissemination of user-generated content, a reliable input backbone is required. Organizational vulnerability to data breaches and other risks is reduced when all data can be kept securely.

Many innovations have occurred s in the field of technology over the past few years which include social networking, Internet of Things, cloud computing and block chain among other technological innovations in information technology. Due to these technological developments, a massive increase in data has been witnessed, and the accumulation of this data has resulted in what is being termed as big data. However, this data may not be beneficial to an organization if it is not analyzed appropriately; therefore inappropriate data is considered as unwanted data. Thus, quality data can be obtained, facilitating proper decision making within the company and bringing focus within the organization.

Therefore, this will direct employees towards the realization of the vision, mission, and objectives of that company. Data quality management (DQM) describes the set of processes which are carried that to maintain high standards of information. These processes range from obtaining the data and implementing innovative data processing techniques in data warehouses to ensure that quality data is acquired. It is acceptable that DQM is a basic achievement in data analysis because quality data helps in deriving actionable and exact intuitions from the obtained information.

II. LITERATURE REVIEW

Jackson Phiri (2023) In a world where there is constant data generation and processing, the need for an integrated system cannot be overemphasized. Studies have shown that standalone desktop spatial systems are often rigid and inflexible to support multiple data processing or demands from multiple users. The integrated spatial management system was designed to address the above highlighted challenges by bringing enhanced possibilities of utilization of spatial data though improving accessibility, visualization, and processing spatial information. The present work employed a mixed approach of qualitative and quantitative techniques to obtain the desired result. Qualitative data collection tools were used to collect field data required to design the prototype. An overview of the architecture of integrated spatial systems consists of a well distributed database linked to multiple tools and platforms to query both the spatial and non-spatial data.

Jean-Louis Tetchueng (2022) The adoption of Cloud Computing services in everyday business life has grown rapidly in recent years due to the many benefits of this paradigm. The various collaboration tools offered by Cloud Computing have eliminated or reduced the notion of distance between entities of the same company or between different organizations. This has led to an increase in the need to share resources (data and services). Community Cloud environments have thus emerged to facilitate interactions between organizations with identical needs and with specific and high security requirements. However, establishing trust and secure resource sharing relationships is a major challenge in this type of complex and heterogeneous environment. This study proposes a trust assessment model (SeComTrust) based on the Zero Trust cyber security strategy. First, the study introduces a community cloud architecture subdivided into different security domains.

Gulain Mugaruka Buduge (2022) This study deals with the implementation of algorithms and tools for the security of academic data protection in the Democratic Republic of the Congo. It consists principally in implementing two algorithms and two distinct tools to secure data and in this particular case, academic data of higher and university education in the Democratic Republic of the Congo. The design of algorithms meets the approach that any researcher in data encryption must use during the development of a computer system. Briefly, these algorithms are steps to follow to encrypt information in any programming language. These algorithms are based on symmetric and asymmetric encryption, the first one uses Christopher Hill's algorithm, which uses texts in the form of matrices before they are encrypted and RSA as one of the asymmetric algorithms, it uses the prime numbers that we have encoded on more than 512 bits. As for tools, we have developed them in php which is only a programming language taken as an example because it is impossible to use all of them.

Grâce Y. E. Johnson (2022) The collaboration tools offered by Cloud Computing have increased the need to share data and services within companies or between autonomous organizations. This has led to the deployment of community cloud infrastructures. However, several challenges will arise from this grouping of heterogeneous organizations. One of the main challenges is the management of trust between the actors of the community. Trust issues arise from the uncertainty about the quality of the resources and entities involved. The quality of a resource can be examined from a security or functional perspective. Therefore, ensuring security and monitoring the quality of resources is to ensure a high level of trust. Therefore, we propose in this study a technique for dynamic trust management and quality monitoring of resources shared between organizations.

Adeyemi A. Adekoya (2020) The purpose of this study is to examine the nature and content of the rapidly evolving undergraduate Principles of Information/Cybersecurity course which has been attracting an ever-growing attention in the computing discipline, for the past decade. More specifically, it is to provide an impetus for the design of standardized principles of Information/Cybersecurity course. To achieve this, a survey of colleges and universities that offer the course was conducted. Several schools of engineering and business, in universities and colleges across several countries were surveyed to generate necessary data. Effort was made to direct the questionnaire only to Computer Information System (CIS), Computer Science (CS), Management Information System (MIS), Information System (IS) and other computer-related departments.

Big Data

Massive data warehouses were required in today's data-driven culture. The widespread use of computers in formerly manual tasks and the proliferation of global online social support networks have both led to the emergence of new phenomena in the management of data produced by such technology. Big data's strength comes from the fact that it can integrate disparate data sources and formats, including but not limited to text, audio, video, images, emails, and sensor readings. There is a chance that these figures are multi- or even unstructured in nature due to the links between machines and humans. The 5Vs and tangible results are what set big data organization apart from data warehousing solutions.

Big Data platforms

Because a listing-based abstraction is built on top of a very large byte-sequential document abstraction, there is impedance mismatch because regular operating systems and their document programmers don't cater to the needs of database programmers. Even though it is typically highly unnatural and may result in poorer performance, covering a high-speed data language with a two-unary-operation run-time like Map reduce is not practical. Hadoop and HDFS are used for big data analytics by many organizations, including major internet businesses. Non-intrusive Map reduction programming approaches are wearing thin with data analysts. They are now deciding between a number of up-to-date languages and frameworks that facilitate the extraction of data diagnostics and the creation and debugging of code. If the data were arbitrarily divided into cubes, then files pertaining to each query might be made accessible.

Public Cloud

Because of cloud computing, the underlying infrastructure may be quickly accessed whenever it's needed. The price of additional configuration options and growth may outweigh the initial expenditure required to launch a cloud-based business. Product cloud providers often market their services by highlighting how simple they are to install for organizations of all sizes. Scalability is typically considered a desirable trait in an organization. However, there are a few dangers associated with using these clouds. Some of the key characteristics of a network cloud are outlined below.

Private Cloud

Specifically, it was managed professionally. Work on the project might be handled internally or by an outside firm under contract. Companies that worry about losing data access may find relief in cloud storage. If your company has difficulties or handles sensitive data, cloud computing may be the way to go. The following are only some of the numerous advantages of CC. get aid and acquire access to water, which is often designated for commercial use. The cloud computing aspect of this service is clear. This implies that companies may look for reliable guidance when deciding which cloud service to use. One such variation is the company-managed private cloud. People living in the cloud wouldn't be able to tell whether their expansion was due to a certain organization.

III. RESEARCH METHODOLOGY

The stream is a place where data frames and processing techniques for research are recorded and limited. In the realm of Big Data processing, the Map Reduce approach has become quite popular. To reduce execution times, stage sizes, and system congestion, Map Reduce is often used. The "management" component of Hadoop is known as "Map Reduce." Map Reduce was created to efficiently handle massive amounts of unstructured work. Compressing files has two key advantages: smaller file sizes and quicker network and disc transfers. Compressing the input might lengthen the time it takes to evaluate data from HDFS. Successful completion of tasks is facilitated by such time management. The architecture explains the general functioning of Big Data processing in distributed programming frameworks. In this context, there might be security issues that arise due to rogue worker nodes or malicious attacks on Map Reduce programming infrastructure. As can be seen from large volumes of data are being processed. In order to secure the infrastructure of Big Data systems, distributed computations and data stores must be secured. To secure the data itself, information dissemination must be privacy-preserving, and sensitive data must be protected through the use of cryptography and other means.

IV. RESULTS

All of the methods save the sixth one are tried out with the aid of the Cladism simulator. The virtual machine monitor (VMM) or hypervisor (HVM) is used to create virtual machines, which may subsequently make requests for hardware resources as required. The Virtual Machine Monitor and a virtual machine may exchange data. To achieve our ends, we plant malicious code in the virtual machines' communication channels. A VM that is behaving abnormally may make excessive requests of the VMM, such as requiring more resources than it requires.

We have included several different sources into the development of the Virtual Cloud. The results are analyzed to show how the inputs to the Cloudsim model led to the expected results. Installing an intrusion detection or prevention system into the virtual machines themselves is one way to increase the safety of a cloud virtualization environment.

Table 1: Overall Response time

Scenario (1-4) SS, IDS/IPS are absent	**Overall Response Time (in milliseconds)**	**VMM Processing Time (in milliseconds)**	**VM Request servicing Time in milliseconds)**	**Memory consumed (in MB) 512MB**
When all VMs are normal	300.382	0.35	0.46	409
When VM1 is infected	315.623	0.49	0.69	427.5
When VM1 infected & affects VM4	329.649	0.51	0.72	445.75
When all VMs infected	395.806	1.21	1.63	449.25

In this initial scenario (5-8), all digital machines are functioning normally in terms of reaction times, memory consumptions (both allocated and absorbed), and support instances. This graph examines the four most recent events for response and service times. This is because each service call from a virtual server requires more time for the IDS/IPS and SS system to execute security checks. The following describes the consequences, assuming Virtual Machine 1 has been breached.

Table 2: VMM Hourly average Response time

Scenario (5-8) SS, IDS/IPS are present	**Overall Response Time(in mille seconds)**	**VMM Processing Time (in mille seconds)**	**VM Request servicing Time in mille seconds)**	**Memory consumed (in MB) 512MB**
When all VMs are normal	348.372	0.32	0.47	409.25
When VM1 is infected	359.219	0.33	0.481	426.25
When VM1 infected & affects VM3 (VM1is stopped processing)	278.256	0.31	0.46	436.66

When all VMs infected (VM3 is stopped processing)	296.15	0.29	0.53	442.33

Here the consequences might go one of eight ways. "Duplication," the most significant new approach, is presented. We utilize a way in which a virtual case is created for each virtual system and then sent off to a pocket virtual machine to perform petition processing instead of having the digital machines handle it directly. More or less instances of a certain virtual server might be determined based on the results of such a calculation. Only one scenario per virtual server is being created for this configuration. By creating instances for each virtual computer, we can see that modern digital machines demand more memory than their predecessors did.

Table 3: VMM Hourly average Response time

Scenario (9-12) with Duplication of 1 instance SS, IDS/IPS are absent	**Overall Response Time (in mille seconds)**	**VMM Processing Time (in mille seconds)**	**VM Request servicing Time in mille seconds)**	**Memory consumed (in MB) 1024MB**
When all VMs are normal	330.492	0.34	0.52	610
When VM1's instance is infected	345.469	0.39	0.47	703.5
When VM1's instance is infected & affects VM2's instance	383.597	0.41	0.52	772.75
When all instances of all VMs infected	480.226	0.62	0.78	881

To improve fault tolerance in certain cases (9-12), we are thinking of running two instances per Virtual machine. The same work may be repeated without interruption using the next case if anything goes wrong with the current one during processing (due to malwares, viruses, or hazards). In all, we have nine cases: seven in "normal" conditions, one corrupted digital machine, and one engineered to keep operating after the first one failed. The identical infection is shown on two different digital computers, and their respective fixes are supplied. We also investigated edge situations involving the other four virtual machines, which were mostly disregarded in the preceding example 2. Our memory use has grown substantially from previous levels since we now duplicate each virtual machine.

V. CONCLUSION

This study explores some of the vulnerabilities inherent to cloud virtualization. Several other strategies were put out by us in this regard. The suggested technology may be used to machine tools, which improves the security of surgical procedures. It is capable of distinguishing between trustworthy machines and those that pose a security risk. IT infrastructures should invest in bonded virtualization technologies due to the variations between real and virtual implementation environments. A high level of security may be provided by an assessment system if neither costs nor problems are involved. The only person who hasn't replicated our methods is the one with the simulator tool. Once a digital computer has been infiltrated, it may infect other digital devices, use vast amounts of memory, and try to obtain access to even more resources. We have a copying mechanism in the fifth and fourth processes that lets us make copies of digital devices and use them in their place; if something were to happen to one of these instances, the other could use the afflicted instance without any interruption in service. VM petition servicing time in milliseconds and memory use, as well as additional data like Total Answer duration, is shown. Our study began with an examination of these issues from the standpoint of the inner workings of the software that implements the virtual machine.

REFERENCES

1. Jean-Louis Tetchueng (2022),"Trust Assessment Model Based on a Zero Trust Strategy in a Community Cloud Environment",Engineering,ISSNno:1947-394X,Vol.14,No.11,Pages.479-496.
2. Grâce Y. E. Johnson (2022),"Shared Resource Quality Monitoring and Dynamic Trust Management in a Community Cloud",Open Journal of Applied Sciences,ISSNno:2165-3925,Vol.12,No.11,Pages.1898-1914.
3. Gulain Mugaruka Buduge (2022),"Algorithms and Tools for Securing and Protecting Academic Data in the Democratic Republic of the Congo",Journal of Information Security,ISSNno:2153-1242,Vol.13,No.4,Pages.312-322.
4. Adeyemi A. Adekoya (2020),"Empirical Evidence for a Descriptive Model of Principles of Information Security Course",Journal of Information Security, ISSNno:2153-1242,Vol.11,No.4,Pages.177-188.
5. Carlos Guillen Gestoso (2022),"LEAN Management Tools Implementation for the Impact Study in the European Foundation for Quality Management Model Score",Open Journal of Business and Management,ISSNno:2329-3292,Vol. 10,No.1,Pages.39-56.
6. Jackson Phiri (2023),"Development of an Integrated System to Enhance Spatial Data Processing & Management for Planning Authorities",Journal of Geoscience and Environment Protection,ISSNno:2327-4344,Vol.11,No.3, Pages.17-29.
7. Bernadette Sharp (2017),"Application of the SECI Model Using Web Tools to Support Diabetes Self-Management and Education in the Kingdom of Saudi Arabia",Intelligent Information Management,ISSNno:2160-5920,Vol.9,No.5, Pages.156-176.
8. Dora Mojoko Ewule (2020),"Community Forest Management: A Strategy for Rehabilitation, Conservation and Livelihood Sustainability: The Case of Mount Oku, Cameroon",Journal of Geoscience and Environment Protection, ISSNno:2327-4344,Vol.8,No.2,Pages.1-14.
9. Md. Kamrul Hasan Chy (2023),"Role of Data Visualization in Finance", American Journal of Industrial and Business Management,ISSNno:2164-5175,Vol.13,No.8,Pages.841-856.
10. Hessa Al Ali (2021),"Implementation Challenges of Data Quality Management—Cases from UAE Public Sector",iBusiness,ISSNno:2150-4083, Vol.13,No.3,Pages.144-153.