Study of Securing in Cloud, Virtual and Big Data Infrastructure

Harsh Arora*

Abstract

There are so many opportunities available in the era of "Big Data" in the fields of advance science, health care improvement, economic growth, social interaction, educational system and many more. Along with big data, another co-coordinating and concerned current topic is cloud computing with virtual machine concept also known as cloud virtual infrastructure. Although from one side there are great benefits from cloud computing and from the another side it enables some threats. More important, these threats are due to cloud virtual infrastructure complexity. It is very important to secure each component in cloud virtual infrastructure which affects the overall system security. This paper explores the security problem in cloud, virtual and big data infrastructure that further identifies security threats and complexities of this virtual infrastructure.

Keywords: Cloud virtual infrastructure, Big data security, Cloud computing, Virtualization security

Introduction

We know big data systems can store very large amounts of data and manage data across many systems. It supports all facilities required for data queries, data consistency and system co-ordination & management system. Big data & cloud data infrastructure are very helpful for the organizations in order to take decisions that are more informed and cost effective. Along with that in cloud computing, improvement in processors & virtual machines technology enhances the data storage processing & related concepts of big data storage but on another hand it is very important to secure cloud, virtual & big data infrastructure. Because in the environment of cloud & big data based software as a service (SaaS) applications & virtualization infrastructure, data integration & co-ordination can become even more complicated because it is huge amount of data where data integrity and data security becomes the most important issue. As we know there is bulk of data present in big data system, so data security criteria has to consider managing and securing structured & unstructured data. However, improvement in big data analytics has presented tools to extract, retrieve, implement, and utilize this data from the complex infrastructure & making violations of privacy easier. Hence, in order to manage the complete database system, the architects have to ensure that complete protection and securing it from unauthorized accessing. In addition, failure to data security and quality reduces the efficiency of criteria of good decision making. Therefore, it is very important to raise security awareness & measures in cloud, virtual & big data environment & infrastructure.

In this paper author has tried to explain the combination of cloud and big data infrastructure, security in cloud, virtual and big data environment, virtualized cloud data infrastructure security, key research challenges in cloud and big data environment.

Harsh Arora*
IINTM, GGSIPU

Cloud & Big Data Infrastructure– "A Powerful Combination"

Most of the organizations are considering moving their big data analytics to one or more cloud delivery models concepts. It is very beneficial for the organizations in today's scenario to use cloud delivery models because they offer exceptional flexibility, enabling it to evaluate the best approach to each business user's request. For example, organizations that already support an internal private cloud environment can add big data analytics

to their in-house offerings, use a cloud services provider, or build a hybrid cloud that protects certain sensitive data in a private cloud but takes advantage of valuable external data sources and applications provided in public clouds.

It is significant to use cloud infrastructure in order to analyze the big data because [8]:

Investments in Big Data Analysis Can Be Significant and Drive a Need for Efficient, Cost-Effective Infrastructure

The resources to support distributed computing models in-house typically reside in large and midsize data centers. Private clouds can offer a more efficient, cost-effective model to implement analysis of big data in-house, while augmenting internal resources with public cloud services. This hybrid cloud option enables companies to use on demand storage space and computing power via public cloud services for certain analytics initiatives (for example, short term projects) and provide added capacity and scale as needed.

Big Data May Mix Internal and External Sources

While enterprises often keep their most sensitive data in-house, huge volumes of big data (owned by the organization or generated party and public providers) may be located externally-some of it already in a cloud environment. Moving relevant data sources behind your firewall can be a significant commitment of resources. Analyzing the data where it resides-either in internal or public cloud data centers or in edge systems and client devices-often makes more sense.

Data Services are needed to Extract Value from Big Data

Depending on requirements and the usage scenario, the best use of IT budget may be to focus on analytics as a service (Aaas)-supported by internal private cloud, a public cloud or a hybrid model.

Security in Cloud, Virtual and Big Data Environment

Big data system contains bulk of data, in that case organizations are facing significant risks and threats to the repositories containing this data, they are just starting down the path of implementing a big data environment as a fact they don't know actually in detail which type of data (structured or unstructured) they want to include in big data repository. So first thing regarding big data is that data should be structured and second thing is it should be secured and the most important criteria is it should be preferred to store the data in various types of clouds depending upon the type of data. The confidence in decision making reduces because of data quality and security failure. So along with the good quality, and good structure of data, it has to be secured as well in order to take decisions correctly and on time.

Organizations should follow a life cycle approach shown in Fig. 1 to examine and extract data across the organization.

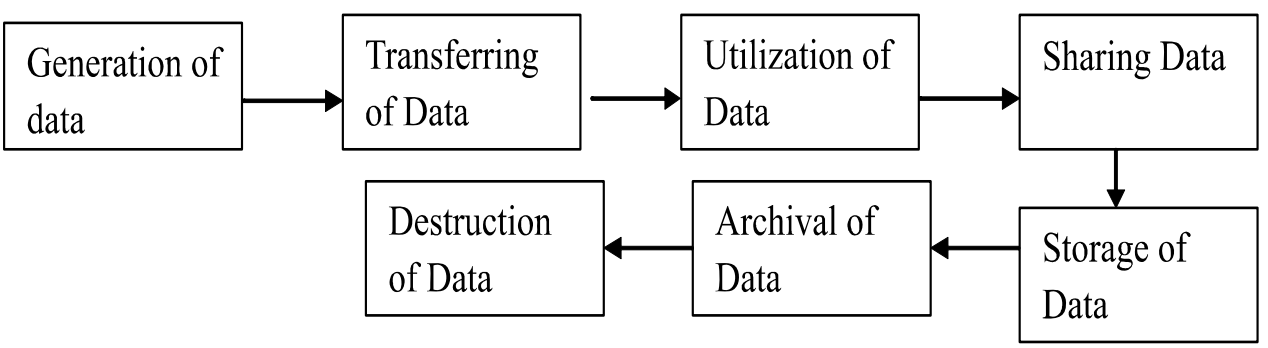

Fig. 1: Life cycle of Data extraction and Data usage in big data environment

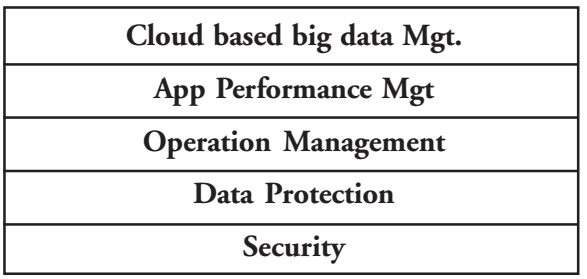

Fig. 2: Data Security Stack

In this life cycle of big data environment, initially the generation of data takes place and then transfering of datafollowed by data utilization. After data utilization, various processes share it. Finally, this data is stored and at the end, irrelevant data is destroyed.

Further, Big data environments should include basic mechanism of control, security and privacy of data as a way to defend and protect the data. As it must be the utmost criteria of data being stored in big data as well as in cloud data storage where three aspects of information security need to be taken care, which are confidentiality, integrity and availability. Out of these three now the concerned topic is confidentiality of data stored in clouds or big data security features. The basic solution for data confidentiality is data encryption, algorithm and key strength. Figure 2 depicts the data security stack.

Proper handling of data and efficient encryption is required to secure transmission. For structured and symmetric data (which should be organized and secured properly in systematic manner for easy storage and extraction) symmetric encryption algorithm is more suitable than asymmetric encryption algorithm. Therefore, organizations can secure big data using data abstraction techniques i.e. masking or encryption.

Virtualized Cloud Data Infrastructure Security

Cloud computing model provides organizations with a more efficient, flexible and cost effective alternative to own their computing resources. However, hackers and security researchers have shown that these capabilities of virtualization can be exploited to create new and more robust forms of malware that are hard to detect and can evade current security technologies [5].

Threat Model

Security responsibility in the cloud is not a single-side responsibility. The cloud provider and the cloud user share it. Customers are not aware of how their Virtual Model (VM) are being protected. On the other hand, the cloud providers running VMs are not aware of the VM contents. Thus, there is no complete trust relationship between cloud customers and providers. In threat model, a hacker can be cloud user that hosts a service or non-cloud user, and in both models, the victim is the cloud provider that runs the service or the other hosted VMs. In the former threat model, hackers have more chances of success, because they have access to the Virtual Cloud Infrastructure (VCI), and can run different malware to gain more access privileges.

Security Threats

Breaching the security of any component in the VCI affects significantly on the security of the other components and consequently affects the overall system security. In these papers [1, 3, 7], the authors investigated different vulnerabilities and security threats in cloud computing focusing on the VCI security threats. Security threats for the cloud virtual infrastructure can be divided into three categories:

Hypervisor Attacks

Hackers consider the hypervisor a potential target because of the greater Control afforded by lower

layers in the system .Compromising the hypervisor enables gaining Control over the installed VMs, the physical system and hosted applications. Hyper Jacking, BLUEPILL, Vitriol, SubVir and DKSM are well-known attacks that target the virtual layer at run-time. These VM-Based Root kits (VMBRs) are capable of inserting a malicious hypervisor *on the fly* or modifying the installed hypervisor to gain control over the host workload. In some hypervisors like Xen, the hypervisor is not alone in administering the VMs. A special privileged VM serves as an administrative interface to Xen, and control the other VMs. This VM is also a potential target for hackers target to exploit vulnerabilities inside that VM to gain access to the hypervisor or the other installed VMs.

VSwitch Attacks

The vSwitch is vulnerable to awide range of layer-2 attacks like a physical switch. These attacks include vSwitch configurations, VLANs and trust zones, and ARP tables [6].

Virtual Machine Attacks

Cloud servers contain tens of VMs, these VMs may be active or offline, and in both states, they are vulnerable to various attacks. Active VMs are vulnerable to all traditional attacks that can affect physical servers. Once a VM is compromised, this gives the VMs on the same physical server a possibility of being able to attack each other, because the VMs share the same hardware and software resources e.g. memory, device drivers, storage, hypervisor software. Colocation of multiple VMs in a single server and sharing the same resources increases the attack surface and the risk of VM-to-VM or VM-to hypervisor compromise. On the other hand, when a physical server is off, it is safe from attacks. However, with VMs when a VM becomes offline, it is still available as VM image files that are susceptible to malware infections and patching. Additionally, provisioning tools and VM templates are exposed to different attacks that target to create new unauthorized VMs, or patch the VM templates to infect the other VMs that will be cloned from this template. These new categories of security threats are a result of the new, complex and dynamic nature of the cloud virtual infrastructure, as follows:

- Multi-Tenancy - Different users within a cloud share the same applications and the physical hardware to run their VMs. This sharing can enable information leakage exploitation and increases the attack surface and the risk of VM-to-VM or VM-to hypervisor compromise.
- Workload Complexity - Server aggregation duplicate the amount of workload and network traffic that runs inside the cloud physical servers, which increase the complexity of managing the cloud workload.
- Loss of Control - users are not aware of the location of their data and services and the cloud providers run VMs they are not aware of their contents.
- Network Topology - The cloud architecture is very dynamic and the existing workload change over time, because of creating and removing VMs. In addition, the mobile nature of the VMs that allowsVMs to migrate from one server to another leads to non-predefined network topology.
- No Physical Endpoints - Due to server and network virtualization, the number of physical endpoints (e.g. switches, servers, NICs) is reduced. These physical endpoints are traditionally used in defining, managing and protecting IT assets.
- Single Point of Access - virtualized servers have a limited number of access points (NICs) available to all VMs. This represents a critical security vulnerability where compromising these access points opens the door to compromise the VCI including VMs, hypervisor or the vSwitch.

Key Research Challenges in Cloud & Big Data Environment

There are many opportunities while working in clouds & big data environment but the question arises on big

data security. So organizations must come to terms with security challenges which can be

- High volume & bulk of data (Nature of big data) reduces the data integrity capability.
- Presence of sensitive data in repository
- Difficulty in making access controls because of aggregated data from multiple sources & unstructured and complicated schema less data from distributed environments.

As we know cloud virtual structure & environment is very complicated, aggregated & dynamic. The virtual architecture of the cloud although provides easy structure of data accessibility from user point of view in distributed environment across the world yet it erases most of the physical boundaries which are normally used in managing & defending the organizational assets leading to complex infrastructure. As a whole there can be following key challenges of big data security [2]:

Internal Clouds are not Inherently Secure

In the past year, many organizations have foregone using public clouds, choosing instead to build private clouds behind their firewalls. This may be the best solution for risk-averse groups. These teams, though, need to understand that just because they have built a cloud inside their firewall does not mean that their solution is safe. It still takes just one bad apple to spoil the barrel—a single department, user or application that is not behaving as it should. An organization that is risk-averse enough to avoid the public cloud should be building a secure cloud—possibly the company should be building its dream cloud, which contains all the security controls that it thinks are missing from a public environment. Since the company physically owns the private cloud, incident response can be very swift. Detection capabilities need to be cloud-specific (for example, sensors need to monitor inside the cloud, not just at its perimeter) and operational capabilities such as patch management must be sharp. A vulnerable service that's in a cloud might have greater exposure and risk than the same service in a standard server farm thanks to the shared nature of cloud resources.

Companies Lack Security Visibility and Risk Awareness

The paucity of security visibility that most providers offer their customers is itself getting plenty of visibility. Obviously, when using a public cloud service, companies must balance the competing factors of control, visibility and cost. This can be a significant issue—reduced visibility results in diminished situational awareness and a questionable understanding of risk. When planning a move to the cloud, an organization needs to recognize this lack of visibility and determine how to best leverage what insight they can get their hands on. Really, this means designing mitigating controls. At the infrastructure and platform levels, this is straightforward: Log more information in your applications and set systems up to generate alerts when signs of compromise or malicious use are spotted (for example, when files are modified, records are changed more frequently than usual, or resource usage is abnormally high). For software as a service (SaaS), though, these precautions will require more thought.SaaS providers are beginning to distinguish themselves via security features. Organizations vetting SaaS providers should consider how they will handle risk awareness—does the provider offer usage data that is granular enough to recognize changes in usage

Sensitive information needs safer storage

Safely storing sensitive information is one of the toughest problems in cloud computing. The solution is to encrypt data, but the critical questions are where to encrypt and how. The first requirement of successful encryption in the cloud, which some providers do not yet understand (or at least don't practice), is: Do not store the encryption key with the encrypted data. Doing so more or less negates any value gained from encrypting the data. However,

the solution is fairly simple, and there's no excuse for not implementing it.In current shared environments, nobody is yet offering a virtual-machine solution that guarantees the integrity of the guest environment. This means that a malicious program could be monitoring the guest's encryption-decryption logic, capturing both plain-text data and the encryption key. Some businesses, though, don't encrypt in the cloud, but encode it before it reaches the cloud service. This works in cases such as a company using a customer resource management system only from its offices, or a business where all users either are at headquarters or VPN into headquarters before connecting to the cloud service.

Apps aren't secure

Application security has been getting attention for years. It's importance increases when an application is deployed to a cloud environment, as the application is more exposed. One of the biggest mistakes an organization can make is to take an existing application and simply deploy it to a cloud without first considering what new attack vectors this move opens up. When possible, an application should be re-architected for cloud deployment—this allows parts of the application to scale independently, and to be more distributed and resilient. It's really an opportunity to make an application more secure than ever. Forcing a development team to not use the corporate firewall as a crutch will result in a solid application. After input and output are taken care of, next up is proper authentication and authorization. These should be checked on every page or service request, not just at initial login. Ideally, any administrative functions are run through a separate application, so if a malicious user does compromise an account, the most he can get is a single user's data, not admin access. The last big thing to consider is data encryption: For performance reasons, most organizations don't want to encrypt all data, so the trick is to find the balance of encrypting enough sensitive information so that if you get compromised, data cannot be pieced together to provide useful identification.

Authentication and authorization must be more robust

Every organization has its own way to manage authentication and authorization. First, it must determine if its current authentication system could also work in a secure and reliable way for users in a cloud environment. If the answer is yes, the follow-up question is whether that is also the best way to authenticate cloud services.

Conclusion

As far as big data, cloud & virtual systems are concerned, there are significant opportunities and dynamic mechanism available which can be availed by multiple users across the world through internet under the umbrella of cloud, virtual & big data infrastructure & environment where multiple data from multiple sources are provided to user but along with that the objective of this paper is to focus on the security measures & solutions of various challenges that are becoming hindrances and obstacles in the path of virtualized cloud environment. The purpose is focusing on developing new virtualization aware security solution that can meet the research challenges concerned with cyber security in big data virtualized clouded environment.

References

1. Bernd Grobauer, Tobias Walloschek and Elmar Stocker, "Understanding Cloud-Computing Vulnerabilities," IEEE Security and Privacy, 10 Jun. 2010. IEEE computer Society Digital Library. IEEE Computer Society, pp.1-8.
2. John Kinsella,"5(more)key cloud security issues"www.csoonline.com/article/717307/5-more-key-cloud-security-issues

3. Kai Hwang, Sameer Kulkareni, Yue Hu, "Cloud Security with Virtualized Defense and Reputation-Based Trust Mangement," Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing, 2009, pp.717-722.

4. Kevin Skapinetz, "Virtualisation as a Blackhat Tool," in *Network Security, Elsevier.* 2007, pp. 4-7.

5. Serdar Cabuk, Chris Dalton, Aled Edwards, et al, "A Comparative Study on Secure Network Virtualization," in Technical Report No. HPL-2008-57, HP Labs, 2008,http://www.hpl.hp.com/techreports/2008/HPL-2008-57.pdf, Accessed on June 2010.

6. W. Dawoud, , Takouna, I., Meinel, C., "Infrastructure as a service security: Challenges and solutions," in t*he 7th International Conference on Informatics and Systems*, Cairo, May 2010.

7. www.intel.in/content/dam/www/public/us/en/documents/products-briefs/big-data-cloud-technologies-brief.pdf