International Journal of Latest Trends in Engineering and Technology
Special Issue SACAIM 2016, pp. 558-566
e-ISSN:2278-621X

ANALYSIS OF SECURITY AND PRIVACY ISSUES ASSOCIATED WITH BIG DATA WITH REFERENCE TO CLOUD ENVIRONMENT

Rakesh Kumar B[1], Puneeth Pai[2] , Arpitha B M[3] and Nisha Dimple Dias[4]

Abstract-In many Enterprise such as marketing, finance, health, manufacturing or service, there is generation of high volume of data and these data needs to be analyzed. Thedata arrives for analysis at high velocity and there is a lot of datavariety.Harnessing big data is complicated, maintaining security and privacy of such datais a big challenge. Security and privacy issues are magnified by Velocity,Volume and variety of Big Datasuch as large cloud infrastructures, diversity of data source and formats, streaming the nature of data acquisition and high volume inter-cloud migration .If a security breach occurs to big data,it would result in even more serious legal repercussions and reputational damage than at present.In this new era, many companies are using the technology to store and analyze petabytes of dataabout their company, business and their customers. As a result, information classificationbecomes even more critical. For making big data secure, techniques such as encryption, logging,honeypot detection must be necessary. In many organizations, the deployment of big data forfraud detection is very attractive and useful.In the area of cyber security, such tasks include user authentication, access control, anomaly detection, user monitoring, and protection from insider threat.The use of data for security tasks is however raising major privacy concerns.This paper provides a summary of the analysis of various measures implemented for big data security and privacy in cloud environment
Keywords : Cloud, Velocity, Honeypot, logging

1 INTRODUCTION

Data volume refers to the large amount of data that are generated every second that is around zeta bytes. Storing of such enormous of data is the biggest challenge. Big data is defined on the basis five Vs that is volume, velocity, variety, value and veracity [1].The Data volumes are increasing rapidly so processing such huge amount of data has become very difficult. During recent years, data production rate has been growing exponentially [2]. Many organizations demand efficient solutions to store and analyze these big amount data that are preliminary generated from various sources such as high throughput instruments, sensors or connected devices.

Data is streaming in at infinite speed and must be dealt within frequent period. To react quickly enough to deal with data velocity is a big challenge for most organizations [1]. Various formats of data are Structured, Unstructured text documents, email, video, audio, bank and financial transactions. Many organizations are still struggling to handle varieties of data [1]. Value is an

[1] *AIMIT, St. Aloysius College, Mangalore, Karnatakam, India*
[2] *AIMIT, St. Aloysius College, Mangalore, Karnatakam, India*
[3] *AIMIT, St. Aloysius College, Mangalore, Karnatakam, India*
[4] *AIMIT, St. Aloysius College, Mangalore, Karnatakam, India*

important feature of the data defined by the added-value that the collected data can bring. So processing big data informative value is important [1]. Big Data veracity ensures that the data used are trusted, authentic and protected from unauthorized access and modification. The data must be secured from its collection, processing and storing on protected and trusted storage facilities [1].

Computers generates highvolumes of data that is primarily generated by Internet of Things (IoT), Next-Generation Sequencing (NGS) machines, scientific simulations and many other sources of data which demand efficient architectures for handling the new datasets. In order to cope with this huge amount of information, "Big Data" solutions such as the Google File System (GFS) [3],MapReduce, Apache Hadoop and the Hadoop Distributed File System (HDFS) have been proposed.

During the past few years, NIST formed the big data working group1 as a community with joint members from industry, academia and government with the aim of developing a consensus definition, taxonomies, secure reference architectures, and technology roadmap. It identifies big data characteristics as extensive datasets that are diverse, including structured, semi-structured, and unstructured data from different domains (variety); large orders of magnitude (volume); arriving with fast rate (velocity); change in other characteristics (variability) [4].

The amount of data generated is expected to double every two years, from 2500 exabytes in 2012 to 40,000 exabytes in 2020 [5]. Security and privacy issues are magnified by the volume, variety, and velocity of Big Data. Large-scale cloud infrastructures, diversity of data sources and formats, the streaming nature of data acquisition and high volume inter-cloud migration all create unique security vulnerabilities.

Big Data and cloud computing are two important issues in the recent years, enables computing resources to be provided asInformation Technology services with high efficiency and effectiveness. Now a days big data is one of the most problemsthat researchers try to solve it and focusing their researches over it to get ride the problem of how big data could be handling inthe recent systems and managed with the cloud of computing, and the one of the most important issue is how to gain a perfectsecurity for big data in cloud computing, our paper reviews a Survey of big data with clouds computing security and themechanisms that used to protect and secure also have a privacy for big data with an available clouds.

There are some sensitive information in the context of cloud computing as it encompasses data from a wide range of areas and disciplines. For example, Data concerning health is a one of the example of the type of sensitive information handled in cloud computing environments, and it is obvious that most individuals will want information related to their health to be secure. Hence, with the proliferation of these new cloud technologies in recent times, privacy and data protection requirements have been evolving to protect individuals against surveillance and database disclosure. Some protective legislation such as EU Data Protection Directive (DPD) and the US Health Insurance Portability and Accountability Act (HIPAA) [6], demand privacy preservation for handling personally identifiable information.

A. Need of security in big data-

In marketing and research, many of the businesses uses big data, but may not have the fundamental assets particularly from a security perspective. When a security breach occurs to big data, it would result in very serious legal repercussions and reputational damage than at present. In the present era, many companies are using the technology to store and analyzeexabytes of data about their company, business and their customers. As a result, information classification

becomes even more critical. For making big data secure, techniques such as encryption, logging, honeypot detection must be necessary. In many organizations, the deployment of big data for fraud detection is very attractive and useful. The challenge of detecting and preventing advanced threats and malicious intruders, must be solved using big data style analysis. These techniques help in detecting the threats in the early stages using more sophisticated pattern analysis and analyzing multiple data sources. With the increase in the use of big data in business, many companies are facing challenges with privacy issues. Data privacy is a liability, thus companies must be on privacy defensive. But unlike security, privacy should be considered as an asset, therefore it becomes a selling point for both customers and other stakeholders. There should be a balance between data privacy and national security.

Information privacy and security is oneof most concerned issues for Cloud Computing due to itsopen environment with very limited user side control. Other considerations are that informationprivacy and security challenges in both Cloud Computingand Big Data must be investigated. the privacy andsecurity providing such forum for researchers, anddevelopers to exchange the latest experience, researchideas and development on fundamental issues andapplications about security and privacy issues in cloud andbig data environments [7].

II ISSUES AND CHALLENGES

Cloud computing comes with numerous security issues because it encompasses manytechnologies including networks, databases, operating systems, virtualization, resourcescheduling, transaction management, load balancing, concurrency control and memorymanagement. Hence, security issues of these systems and technologies are applicable to cloudcomputing. For example,it is very important for the network which interconnects the systems in acloud to be secure. Also, virtualization paradigm in cloud computing results in several securityconcerns. For example, mapping of the virtual machines to the physical machines has to be performed very securely.

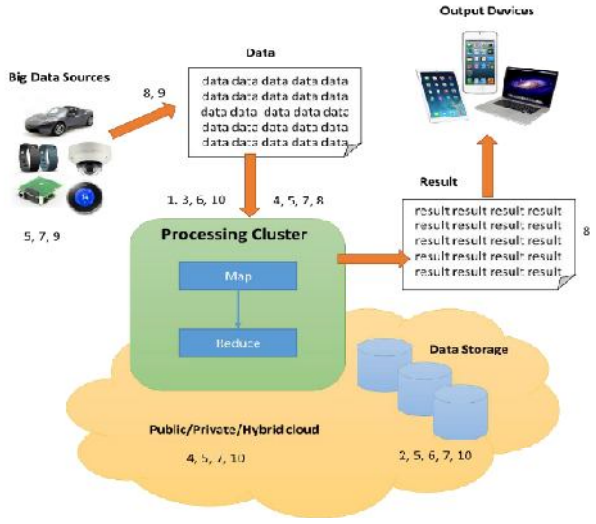

Figure 1: Privacy and Security Challenges in Big Data ecosystem

Data security not only involves the encryption of the data, but also ensures that appropriatepolicies are enforced for data sharing. In addition, resource allocation and memory managementalgorithms also have to be secure. The big data issues are most acutely felt in certain industries,such as telecoms, web marketing and advertising, retail and financial services, and certaingovernment activities. The data explosion is going to make life difficult in many industries, andthe companies will gain considerable advantage which is capable to adapt well and gain theability to analyze such data explosions over those other companies. Finally, data miningtechniques can be used in the malware detection in clouds.

The challengesof security in cloud computing environments can be categorized into networklevel, user authentication level, data level, and genericissues.

Network level:The challenges that can be categorized under a network level deal with networkprotocols and network security, such as distributed nodes, distributed data, Internodecommunication.

***Authentication level*:** The challenges that can be categorized under user authentication level dealswith encryption/decryption techniques, authentication methods such asadministrative rights fornodes, authentication of applications and nodes, andlogging.

Data level :The challenges that can be categorized under data level deals with data integrity andavailability such as data protection and distributed data.

Generic types: The challenges that can be categorized under general level are traditional securitytools, anduse of different technologies

Google has introduced MapReduce[8] framework for processing large amounts of data on commodity hardware. Apache's Hadoop distributed file system (HDFS) is evolving as a superior software component for cloud computing combined along with integrated parts such as MapReduce. Hadoop, which is an open-source implementation of Google MapReduce, including a distributed file system, provides to the application programmer the abstraction of the map and the reduce. With Hadoop it is easier for organizations to get a control on the large volumes of data being generated everyday, but at the same time can also create problems related to security, monitoring, data access,high availability and business continuity.

III REVIEW OF SECRUITY AND PRIVACY MESAURES

Big Data remains one of the most talked about technologytrends in 2013. But lost among all the excitement about thepotential of Big Data are the very real security and privacychallenges that threaten to slow this momentum. Securityand privacy issues are magnified by the three V's of bigdata: Velocity, Volume, and Variety. These factors includevariables such as large-scale cloud infrastructures,diversity of data sources and formats, streaming nature ofdata acquisition and the increasingly high volume of intercloudmigrations. Consequently, traditional securitymechanisms, which are tailored to securing small-scalestatic (as opposed to streaming) data, often fall short.

The CSA's Big Data Working Group followed a three stepprocess to arrive at top security and privacy challengespresented by Big Data; Interviewed CSA members andsurveyed security practitioner oriented trade journals todraft an initial list of high priority security and privacyproblems studied published solutions. Characterized aproblem as a challenge if the proposed solution does notcover the problem scenarios. Following this exercise, theWorking Group researchers compiled their list of the Top10 challenges as shown in figure 2 below.

The information security practitioners at the CloudSecurity Alliance know that big data and analytics systemsare here to stay. They also agree on the big questions thatcome next: How can we make the systems that store andcompute the data secure? And, how can we ensure privatedata stays private as it moves through different stages ofanalysis, input and output? The answers to those questionsthat prompted the group's latest 39-page report detailing
10 major security and privacy challenges facinginfrastructure providers and customers. By outlining theissues involved, along with analysis of internal andexternal threats and summaries of current approaches tomitigating those risks, the alliance's members hope to approach technology vendors, academic researchers andpractitioners to collaborate on computing techniques andbusiness practices that reduce the risks associated with analyzing massive datasets using innovative data analytics.Existing encryption technologies that don't scale well tolarge datasets. Real-time system monitoring techniquesthat works well on smaller volumes of data but not verylarge datasets. The growing number of devices, fromsmartphones to sensors, producing data for analysis.

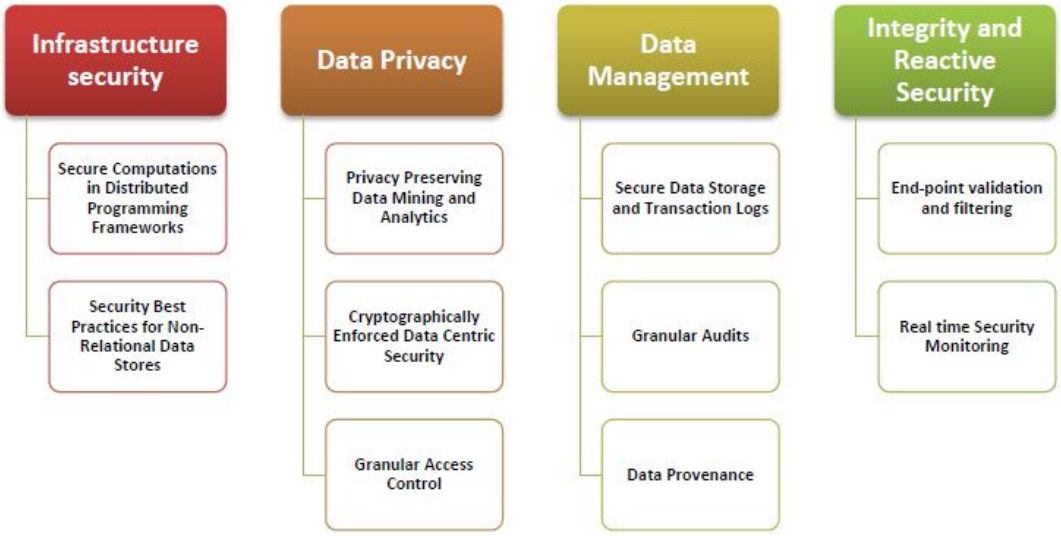

Figure 2: 10 Challenges of CSA's Big Data Working Group

B. Security Issues in Cloud Computing -

a. Multi-tenancy: Multi-tenancy refers to sharing physical devices and virtualized resources between multiple independent users. Using this kind of arrangement means that an attacker could be on the same physical machine as the target. Cloud providers use multi-tenancy features to build infrastructures that can efficiently scale to meet customers' needs, however the sharing of resources means that it can be easier for an attacker to gain access to the target's data.

b. Loss of Control: Loss of control is another potential breach of security that can occur where consumers' data, applications, and resources are hosted at the cloud provider's owned premises. As the users do not have explicit control over their data, this makes it possible for cloud providers to perform data mining over the users' data, which can lead to security issues. In addition, when the cloud providers backup data at different data centers, the consumers cannot be

sure that their data is completely erased everywhere when they delete their data. This has the potential to lead to misuse of the unerased data.

c. Trust Chain in Clouds: Trust plays an important role in attracting more consumers by assuring on cloud providers. Due to loss of control (as discussed earlier), cloud users rely on the cloud providers using trust mechanisms as an alternative to giving users transparent control over their data and cloud resources. Therefore, cloud providers build confidence amongst their customers by assuring them that the provider's operations are certified in compliance with organizational safeguards and standards.

d. Cloud Security otherchallenges

Generic types: The challenges that can be categorized under general level are traditional security tools, anduse of different technologies

e. Distributed Nodes

Distributed nodes [9]are an architectural issue. The computation is done in any set of nodes.Basically, data is processed in those nodes which have the necessary resources. Since it canhappen anywhere across the clusters, it is very difficult to find the exact location of computation.Because of this it is very difficult to ensure the security of the place where computation is done.

f. Distributed Data

In order to alleviate parallel computation,a large data set can be stored in many pieces acrossmany machines. Also, redundant copies of data are made to ensure data reliability. In case aparticular chunk is corrupted, the data can be retrieved from its copies. In the cloud environment,it is extremely difficult to find exactly where pieces of a file are stored. Also, these pieces of dataare copied to another node/machines based on availability and maintenance operations. Intraditional centralized data security system, critical data is wrapped around various security tools.

g. Internode Communication

Much Hadoop distributions use RPC over TCP/IP for user data/operational data transfer betweennodes. This happens over a network, distributed around globe consisting of wireless and wirednetworks. Therefore, anyone can tap and modify the inter node communication [9] for breakinginto systems.

h. Data Protection

Many cloud environments like Hadoop store the data as it is without encryption to improveefficiency. If a hacker can access a set of machines, there is no way to stop him to steal the criticaldata present in those machines.

i. Administrative Rights for Nodes

A node has administrative rights [9] and can access any data. This uncontrolled access to anydata is very dangerous as a malicious node can steal or manipulate critical user data.

j. Authentication of Applications and Nodes

Nodes can join clusters to increase the parallel operations. In case of no authentication, third partnodes can join clusters to steal user data or disrupt the operations of the cluster.

k. Logging

In the absence of logging in a cloud environment, no activity is recorded which modify or deleteuser data. No information is stored like which nodes have joined cluster, which Map

Reduce jobshave run, what changes are made because of these jobs. In the absence of these logs, it is verydifficult to find if someone has breached the cluster if any, malicious altering of data is donewhich needs to be reverted. Also, in the absence of logs, internal users can do malicious datamanipulations without getting caught.

l. Traditional Security Tools

Traditional security tools are designed for traditional systems where scalability is not huge ascloud environment. Because of this, traditional security tools which are developed over yearscannot be directly applied to this distributed form of cloud computing and these tools do not scaleas well as the cloud scales.

m. Use of Different Technologies

Components include database, computing power, network, and many other stuff. Because of thewide use of technologies, a small security weakness in one component can bring down the wholesystem. Because of this diversity, maintaining security in the cloud is very challenging.

Use of SSL/TLS Implement Secure Socket Layer (SSL) or Transport Layer Security (TLS) between nodes, and between nodes and applications. Cloud era offers TLS, and some cloud providers offer secure communication, otherwise user need to integrate these services into their application stack [10].

IV REVIEW OF THE PROPOSED APPROACHES

We present various security measures which would improve the security of cloud computingenvironment. Since the cloud environment is a mixture of many different technologies, wepropose various solutions which collectively will make the environment secure. The proposedsolutions encourage the use of multiple technologies/ tools to mitigate the security problems specified in previous sections. Security recommendations are designed such that they do notdecrease the efficiency and scaling of cloud systems.Following security measures should be taken to ensure the security in a cloud environment.

File Encryption

Since the data is present in the machines in a cluster, a hacker can steal all the criticalinformation. Therefore, all the data stored should be encrypted. Different encryption keys shouldbe used on different machines and the key information should be stored centrally behind strongfirewalls. This way, even if a hacker is able to get the data, he cannot extract meaningfulinformation from it and misuse it. User data will be stored securely in an encrypted manner.

Network Encryption

All the network communication should be encrypted as per industry standards. The RPCprocedure calls which take place should happen over SSL so that even if a hacker can tap intonetwork communication packets, he cannot extract useful information or manipulate packets.

Logging

All the map reduce jobs which modify the data should be logged. Also, the information of users,which are responsible for those jobs should be logged. These logs should be audited regularly tofind if any, malicious operations are performed or any malicious user is manipulating the data inthe nodes.

Software Format and Node Maintenance

Nodes which run the software should be formatted regularly to eliminate any virus present. Allthe application software and Hadoop software should be updated to make the system moresecure.

Nodes Authentication

Whenever a node joins a cluster, it should be authenticated. In case of a malicious node, it shouldnot be allowed to join the cluster. Authentication techniques like Kerberos can be used to validatethe authorized nodes from malicious ones.

Rigorous System Testing of Map Reduce Jobs

After a developer writes a map reduce job, it should be thoroughly tested in a distributedenvironment instead of a single machine to ensure the robustness and stability of the job.

Honeypot Nodes

Honey pot nodes should be present in the cluster, which appear like a regular node but is a trap.These honeypots trap the hackers and necessary actions would be taken to eliminate hackers.

Layered Framework for Assuring Cloud

A layered framework for assuring cloud computing consists of thesecure virtual machine layer, secure cloud storage layer, secure cloud data layer, and the securevirtual network monitor layer. Cross cutting services are rendered by the policy layer, the cloudmonitoring layer, the reliability layer and the risk analysis layer.

Third Party Secure Data Publication to Cloud

Cloud computing helps in storing of data at a remote site inorder to maximize resource utilization.Therefore, it is very important for this data to be protected and access should be given only toauthorized individuals. Hence this fundamentally amounts to secure third party publication ofdata that is required for data outsourcing, as well as for external publications. In the cloudenvironment, the machine serves the role of a third party publisher, which stores the sensitive datain the cloud. This data needs to be protected, and the above discussed techniques have to beapplied to ensure the maintenance of authenticity and completeness.

Access Control

Real time access control will be a good security measure in the cloud environment. In addition toaccess control to the cloud environment, operational control within a database in the cloud can beused to prevent configuration drift and unauthorized application changes. Multiple factors such asIP address, time of the day, and authentication method can be used in a flexible way to employabove measures. For example, access can be restricted to specific middle tier, creating a trustedpath to the data. Keeping a security administrator separate from the database administrator will bea good idea. The label security method will be implemented to protect sensitive data by assigningdata label or classifying data.

Data can be classified as public, confidential and sensitive. If the user label matches with the labelof the data, then access is provided to the user.

V.CONCLUSION

The impact big data has created and will continue to create can ripple through all facets of our life. Global Data is on the rise, by 2020, we would have quadrupled the data we generate every

day. This data would be generated through a wide array of sensors we are continuously incorporating in our lives. Data collection would be aided by what is today dubbed as the "Internet of Things". Big Data does not arise out of a vacuum: it is recorded from some data generating source.

Enterprise data security is a challenging task to implement. In this paper, we have highlighted the top ten security and privacy problems that need to be addressed to make Big Data processing and computing infrastructure more secure and we have also reviewed several security and privacy issues on big data in the cloud. We also discussed several security challenges that are raised by existing or forthcoming privacy legislation, such as the EU DPD and the HIPAA. In this paper, we have highlighted the top ten security and privacy problems that need to be addressed to make Big Data processing and computing infrastructure more secure. We can conclude that no security measure is ultimate there will also be new challenges with the changing technology and time

REFERENCES

[1] A Katal, M Wazid, and RHGoudar , "Big data: Issues, challenges, tools and Good practices",pp404 – 409, 2013.

[2] A. Szalay and J. Gray, "2020 Computing: Science in an exponential world," Nature, vol. 440, pp. 413–414, 2006.

[3 http://www.informationweek.com/big-data/bigdataanalytics/big-databrings-big-security-problems/d/did/1252747.

[4 NIST Special Publication 15001–291 version 1, Definitions and Taxonomies Subgroup, September 2015,

[5] http://www.emc.com/leadership/digital-universe/iview/big-data-2020.html.

[6] U. States., "Health insurance portability and accountability act of 1996 [micro form]: conference report (to accompany h.r. 3103)." http://nla.gov.au/nla.catvn4117366, 1996.

[7] S Ghemawat, H Gobioff and S TLeung , "The Google File System" , SOSP , 2003.

[8] Ren, Yulong, and Wen Tang. "A SERVICE INTEGRITY ASSURANCE FRAMEWORK FOR CLOUD COMPUTING BASED ON MAPREDUCE", Proceedings of IEEE CCIS2012. Hangzhou, pp 240 – 244, 2012.

[9] "Securing Big Data: Security Recommendations forHadoop and NoSQL Environments", *Securosisblog*, version 1.0 (2012).

[10] Securing Big Data: Security Recommendations for Hadoop and NoSQL Environments, Version 1.0 Released: October 12, 2012 , licensed by vormetric