ISSN:0975-9646

K.Shanmugapriya et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (6) , 2015, 4952-4956

Security Issues Associated With Big Data in Cloud Computing

K.Shanmugapriya [1],M.Murugeswari[2]
III MCA,Department of Computer Applications,
Dhanalakshmi Srinivasan College of Arts and Science for Women,
Perambalur.

K.Suriya[3] MCA., M.Phil,
Assistant Professor, Department of Computer Applications,
Dhanalakshmi Srinivasan College of Arts and Science for Women
Perambalur.

Abstract-Big Data and cloud computing are two important issues in the recent years, enables computing resources to be provided as Information Technology services with high efficiency and effectiveness. Now a day's big data is one of the most problems that researchers try to solve it and focusing their researches over it to get ride the problem of how big data could be handling in the recent systems and managed with the cloud of computing, and the one of the most important issue is how to gain a perfect security for big data in cloud computing, In this paper, we discuss security issues for cloud computing, Big data, Map Reduce and Hadoop environment. The main focus is on security issues in cloud computing that are associated with bigdata. Big data applications are a great benefit to organizations, business, companies and many large scale and small scale industries. We also discuss various possible solutions for the issues in cloud computing security and Hadoop. Cloud computing security is developing at a rapid pace which includes computer security, network security, information security, and data privacy.Moreover, cloud computing, big data and its applications, advantages are likely to represent the most promising new frontiers in science.

Keywords
Cloud Computing, Big Data, Hadoop,Map Reduce, HDFS (Hadoop Distributed File System).

1. INTRODUCTION

In order to analyze complex data and to identify patterns it is very important to securely store, manage and share large amounts of complex data.Cloud comes with an explicit security challenge, i.e. the data owner might not have any control of where the data is placed. The reason behind this control issue is that if one wants to get the benefits of cloud computing, he/she must also utilize the allocation of resources and also the scheduling given by the controls. Hence it is required to protect the data in the midst of untrustworthy processes. Since cloud involves extensive complexity, we believe that rather than providing a holistic solution to securing the cloud, it would be ideal to make noteworthy enhancements in securing the cloud that will ultimately provide us with a secure cloud.Fig.1

Google has introduced MapReduce [1] 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 .

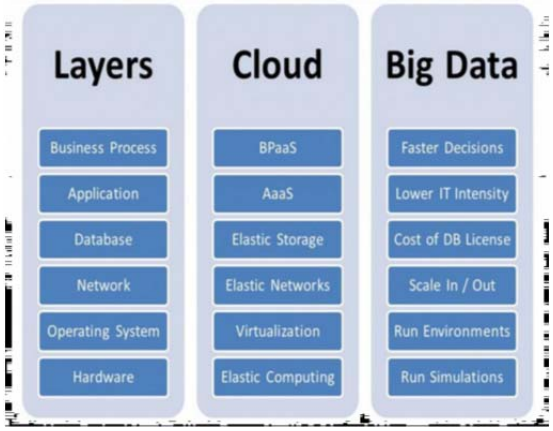

Fig. 1. Big Data and clouds

In this paper, we come up with some approaches in providing security. We ought a system that can scale to handle a large number of sites and also be able to process large and massive amounts of data. However, state of the art systems utilizing HDFS and MapReduce are not quite enough/sufficient because of the fact that they do not provide required security measures to protect sensitive data. Moreover, Hadoop framework is used to solve problems and manage data conveniently by using different techniques such as combining the k-means with data mining technology [3].

1.1 Cloud Computing

Cloud Computing is a technology which depends on sharing of computing resources than having local servers or personal devices to handle the applications. In Cloud Computing, the word "Cloud" means "The Internet", so Cloud Computing means a type of computing in which services are delivered through the Internet. The goal of Cloud Computing is to make use of increasing computing power to execute millions of instructions per second.

Cloud computing technology is being used to minimize the usage cost of computing resources [4]. The cloud network, consisting of a network of computers, handles the load instead. The cost of software and hardware on the user end

decreases. The only thing that must be done at the user's end is to run the cloud interface software to connect to the cloud. Cloud Computing consists of a front end and back end. The front end includes the user's computer and software required to access the cloud network. Back end consists of various computers, servers and database systems that create the cloud.

The user can access applications in the cloud network from anywhere by connecting to the

Cloud using the Internet. Some of the realtime applications which use CloudComputing are Gmail, GoogleCalendar, Google Docs andDropboxetc.,Fig.2

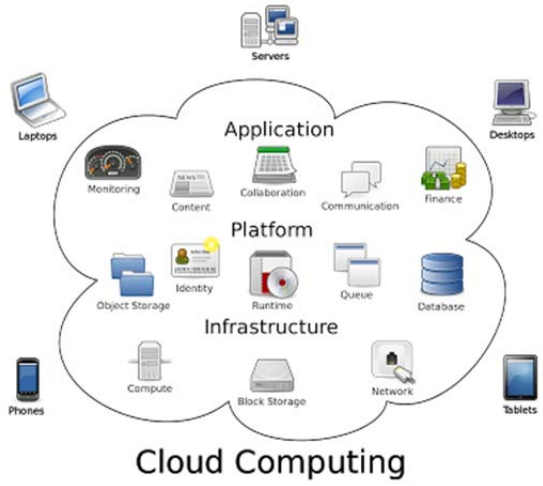

Fig.2 Cloud Computing

1.2 Big Data

Big Data Fig.3 is the word used to describe massive volumes of structured and unstructured data that are so large that it is very difficult to process this data using traditional databases and software technologies. The term "Big Data [5]" is companies who had to query loosely structured very large distributed data. The three main terms that signify Big Data have the following properties:

a) Volume: Many factors contribute towards increasing Volume streaming data and data collected from sensors etc.,

b) Variety: Today data comes in all types of formats emails, video, audio, transactions etc.,

c) Velocity: This means how fast the data is being produced and how fast the data needs to be processed to meet the demand.

Fig.3 Big Data

1.3 Hadoop

Hadoop, which is a free, Java-based programming framework supports the processing of large sets of data in a distributed computing environment. It is a part of theApacheproject sponsored by the Apache Software Foundation. Hadoop cluster uses a Master/Slave structure [6]. Using Hadoop, large data sets can be processed across a cluster of servers and applications can be run on systems with thousands of nodes involving thousands ofterabytes. Distributed file system in Hadoop helps in rapid data transfer rates and allows the system to continue its normal operation even in the case of some node failures. This approach lowers the risk of an entire system failure, even in the case of a significant number of node failures. Hadoop enables a computing solution that is scalable, cost effective, flexible and fault tolerant. Hadoop Framework is used by popular companies like Google, Yahoo, Amazon and IBM etc., to support their applications involving huge amounts of data. Hadoop has two main sub projects – Map Reduce and Hadoop Distributed File System (HDFS).

1.4 Map Reduce

Hadoop Map Reduce is a framework [7] used to write applications that process large amounts of data in parallel on clusters of commodity hardware resources in a reliable, fault-tolerant manner.

A Map Reduce job first divides the data into individual chunks which are processed by Map jobs in parallel. The outputs of the maps sorted by the framework are then input to the reduce tasks. Generally the input and the output of the job are both stored in a file-system. Scheduling, Monitoring and re-executing failed tasks are taken care by the framework.

1.5 Hadoop Distributed File System (HDFS)

HDFS [8] is a file system that spans all the nodes in a Hadoop cluster for data storage. It links together file systems on local nodes to make it into one large file system. HDFS improves reliability by replicating data across multiple sources to overcome node failures.

2. Cloud Computing In Big Data

The rise of cloud computing and cloud data stores has been a precursor and facilitator to the emergence of big data. Cloud computing is the commoditization of computing time and data storage by means of standardized technologies. It has significant advantages over traditional physical deployments. However, cloud platforms come in several forms and sometimes have to be integrated with traditional architectures. This leads to confuse for decision makers in charge of big data projects, leads to a question of how and which cloud computing is the optimal choice for their computing needs, especially if it is a big data project? These projects regularly exhibit unpredictable, bursting, or immense computing power and storage needs. At the same time business stakeholders expect swift, inexpensive, and dependable products and project outcomes.

3.BIGDATA APPLICATIONS

The big data application refers to the large scale distributed applications which usually work with large data sets. Data exploration and analysis turned into a difficult problem in many sectors in the span of big data.With large and complex data, computation becomes difficult to be handled bythe traditional data processing applicationswhich triggers the development of big data applications [9]. Google's map reduce framework and apache Hadoop are the defacto software systems [10] for big data applications, in which these applications generates a huge amount of intermediate data.

Manufacturing and Bioinformatics are the two major areas of big data applications. Big data provide an infrastructure for transparency in manufacturing industry, which has the ability to unravel uncertainties such as inconsistent component performance and availability. In these big data applications, a conceptual framework of predictive manufacturing begins with data acquisition where there is a possibility to acquire different types of sensory data such as pressure, vibration, acoustics, voltage, current, and controller data. The combination of sensory data and historical data constructs the big data in anufacturing. This generated big data from the above combination acts as the input into predictive tools and preventive strategies such as prognostics and health management. Another important application for Hadoop is Bioinformatics which covers the next generation sequencing and other biological domains. Bioinformatics [11] which requires a large scale data analysis, uses Hadoop. Cloud computing gets the parallel distributed computing framework together with computer clusters and web interfaces.

4. ADVANTAGES OF BIGDATA

In Big data, the software packages provide a rich set of tools and options where an individual could map the entire data landscape across the company, thus allowing the individual to analyze the threats he/she faces internally. This is considered as one of the main advantages as big data keeps the data safe. With this an individual can be able to detect the potentially sensitive information that is not protected in an appropriate manner and makes sure it is stored according to the regulatory requirements.

All the organizations and business would benefit from speed, capacity, and scalability of cloud storage. Moreover, end users can visualize the data and companies can find new business opportunities.Another notable advantage with big-data is, data analytics, which allow the individual to personalize the content or look and feel of the website in real time so that it suits the each customer entering the website. If big data are combined with predictive analytics, it produces a challenge for many industries. The combination results in the exploration ofthese four areas:

a) Calculate the risks on large portfolios
b) Detect, prevent, and re-audit financial fraud
c) Improve delinquent collections
d) Execute high value marketing campaigns

5. NEED OF SECURITY IN BIGDATA

For marketing and research, many of the businesses uses big data, but may not have the fundamental assets particularly from a security perspective. If a security breach occurs to big data, it wouldresultin 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 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. Not only security but also data privacy challenges existing industries and federal organizations. With the increase in the use of big data in business, many companies are wrestling 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.

6. ISSUES AND CHALLENGES

Cloud computing comes with numerous security issues because it encompasses many

technologies including networks, databases, operating systems, virtualization, resource scheduling, transaction management, load balancing, concurrency control and memory management. Hence, security issues of these systems and technologies are applicable to cloud computing. For example,it is very important for the network which interconnects the systems in a cloud to be secure. Also, virtualization paradigm in cloud computing results in several security concerns. For example, mapping of the virtual machines to the physical machines has to be performed very securely. Data security not only involves the encryption of the data, but also ensures that appropriate policies are enforced for data sharing. In addition, resource allocation and memory management algorithms 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 certain government activities.

The data explosion is going to make life difficult in many industries, and the companies will gain considerable advantage which is capable to adapt well and gain the ability to analyze such data explosions over those other companies. Finally, data mining techniques can be used in the malware detection in clouds. The challengesof security in cloud computing environments can be categorized into network level, user authentication level, data level, and genericissues.

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

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

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

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

7.THE PROPOSED APPROACHES

We present various security measures which would improve the security of cloud computing environment. Since the cloud environment is a mixture of many different technologies, we propose various solutions which collectively will make the environment secure. The proposed solutions encourage the use of multiple technologies/ tools to mitigate the security problem specified in previous sections. Security recommendations are designed such that they do not decrease the efficiency and scaling of cloud systems. Following security measures should be taken to ensure the security in a cloud environment.

7.1 File Encryption

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

7.2 Network Encryption

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

7.3 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 to find if any, malicious operations are performed or any malicious user is manipulating the data in the nodes.

7.4 Software Format and Node Maintenance

Nodes which run the software should be formatted regularly to eliminate any virus present. All the application softwares and Hadoop software should be updated to make the system more secure.

7.5 Nodes Authentication

Whenever a node joins a cluster, it should be authenticated. In case of a malicious node, it should not be allowed to join the cluster. Authentication techniques like Kerberos can be used to validate the authorized nodes from malicious ones[13].

7.6 Rigorous System Testing of Map Reduce Jobs

After a developer writes a map reduce job, it should be thoroughly tested in a distributed
environment instead of a single machine to ensure the robustness and stability of the job.[14]

7.7 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.

7.8 Layered Framework for Assuring Cloud

A layered framework for assuring cloud computing [15] as shown in Figure (1) consists of thesecure virtual machine layer, secure cloud storage layer, secure cloud data layer, and the secure virtual network monitor layer. Cross cutting services are rendered by the policy layer, the cloud monitoring layer, the reliability layer and the risk analysis layer.

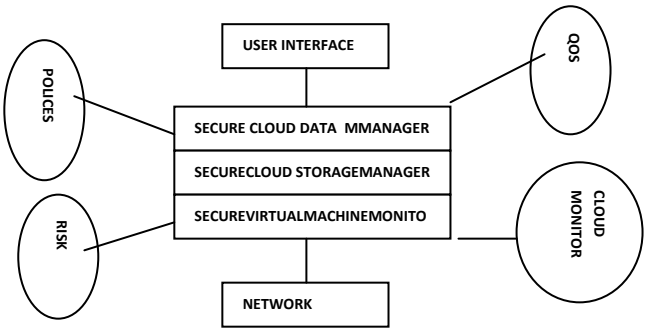

Fig4: Layered framework for assuring cloud [15]

7.9 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 to authorized individuals. Hence this fundamentally amounts to secure third party publication of data that is required for data outsourcing, as well as for external publications. In the cloud environment, the machine serves the role of a third party publisher, which stores the sensitive data in the cloud. This data needs to be protected, and the above discussed techniques have to be applied to ensure the maintenance of authenticity and completeness.

7.10 Access Control

Integration of mandatory access control and differential privacy in distributed environment will be a good security measure. Data providers will control the security policy of their sensitive data. They will also control the mathematical bound on privacy violation that could take place. In the above approach, users can perform data computation without any leakage of data. To prevent information leak, SELinux [16] will be used. SELinux is nothing but Security-Enhanced Linux, which is a feature that provides the mechanism for supporting access control security

policy through the use of Linux Security Modules (LSM) in the Linux Kernel. Enforcement of differential privacy will be done using modification to Java Virtual Machine and the Map Reduce framework. It will have inbuilt applications which store the user identity pool for the whole cloud service. So the cloud service will not have to maintain each user's identity for each application. In addition to the above methodologies, cloud service will support third party authentication. The third party will be trusted by both the cloud service and accessing user. Third party authentication will add an additional security layer to the cloud service.

Real time access control will be a good security measure in the cloud environment. In addition to access control to the cloud environment, operational control within a database in the cloud can be used to prevent configuration drift and unauthorized application changes. Multiple factors such as IP address, time of the day, and authentication method can be used in a flexible way to employ above measures. For example, access can be restricted to specific middle tier, creating a trusted path to the data. Keeping a security administrator separate from the database administrator will be a good idea. The label security method will be implemented to protect sensitive data by assigning data label or classifying data. Data can be classified as public, confidential and sensitive. If the user label matches with the label of the data, then access is provided to the user. Examination of numerous data breaches has shown that auditing could have helped in early detection of problems and avoids them. Auditing of events and tracking of logs taking place in the cloud environment will enable possible attack.

8. CONCLUSION

Cloud environment is widely used in industry and research aspects; therefore security is animportant aspect for organizations running on these cloud environments. Using proposed approaches, cloud environments can be secured for complex business operations.

REFERENCES

[1] Ren, Yulong, and Wen Tang. "A SERVICE INTEGRITY ASSURANCE FRAMEWORK FOR CLOUD COMPUTING BASED ON MAPREDUCE."*Proceedings of IEEE CCIS2012*. Hangzhou: 2012, pp 240 – 244, Oct. 30 2012-Nov. 1 2012

[2] N, Gonzalez, Miers C, Redigolo F, Carvalho T, Simplicio M, de Sousa G.T, and Pourzandi M. "A Quantitative Analysis of Current Security Concerns and Solutions for Cloud Computing.". Athens: 2011., pp 231 – 238, Nov. 29 2011- Dec. 1 2011

[3] Hao, Chen, and Ying Qiao. "Research of Cloud Computing based on the Hadoop platform.". Chengdu, China: 2011, pp. 181 – 184, 21-23 Oct 2011.

[4] Y, Amanatullah, Ipung H.P., Juliandri A, and Lim C. "Toward cloud computing reference architecture: Cloud service management perspective.". Jakarta: 2013, pp. 1-4, 13-14 Jun. 2013.

[5] A, Katal, Wazid M, and Goudar R.H. "Big data: Issues, challenges, tools and Good practices.". Noida: 2013, pp. 404 – 409, 8-10 Aug. 2013.

[6] Lu, Huang, Ting-tin Hu, and Hai-shan Chen. "Research on Hadoop Cloud Computing Model and its Applications.". Hangzhou, China: 2012, pp. 59 – 63, 21-24 Oct. 2012.

[7] Wie, Jiang , Ravi V.T, and Agrawal G. "A Map-Reduce System with an Alternate API for Multi-core Environments.". Melbourne, VIC: 2010, pp. 84-93, 17-20 May. 2010.

[8] K, Chitharanjan, and Kala Karun A. "A review on hadoop — HDFS infrastructure extensions.". JeJu Island: 2013, pp. 132-137, 11-12 Apr. 2013.

[9] F.C.P, Muhtaroglu, Demir S, Obali M, and Girgin C. "Business model canvas perspective on big data applications." *Big Data, 2013 IEEE International Conference*, Silicon Valley, CA, Oct 6-9, 2013,

[10] Zhao, Yaxiong , and Jie Wu. "Dache: A data aware caching for big-data applications using the MapReduce framework." *INFOCOM, 2013 Proceedings IEEE*, Turin, Apr 14-19, 2013, pp. 35 - 39.

[11] Xu-bin, LI , JIANG Wen-rui, JIANG Yi, ZOU Quan "Hadoop Applications in Bioinformatics." *Open Cirrus Summit (OCS), 2012 Seventh*, Beijing, Jun 19-20, 2012, pp. 48 - 52.

[12] Bertino, Elisa, Silvana Castano, Elena Ferrari, and Marco Mesiti. "Specifying and enforcing access control policies for XML document sources." pp 139-151.

[13] E, Bertino, Carminati B, Ferrari E, Gupta A , and Thuraisingham B. "Selective and Authentic Third- Party Distribution of XML Documents."2004, pp. 1263 - 1278.

[14] Kilzer, Ann, Emmett Witchel, Indrajit Roy, Vitaly Shmatikov, and Srinath T.V. Setty. "Airavat:Security and Privacy for MapReduce."

[15] P.R , Anisha, Kishor Kumar Reddy C, Srinivasulu Reddy K, and Surender Reddy S. "Third Party Data Protection Applied To Cloud and Xacml Implementation in the Hadoop Environment With Sparql."2012. 39-46, Jul – Aug. 2012.

[16] "Security-Enhanced Linux."*Security-Enhanced Linux*. N.p. Web. 13 Dec 2013.