INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 2 ISSUE 6 (2016) PAGES 1462 - 1466
RECEIVED : 24/11/2016. PUBLISHED: 08/12/2016
December 8, 2016

SECURITY ISSUES ASSOCIATED WITH BIG DATA IN CLOUD COMPUTING

[1]Sandrilla.R, [2]Anjugam.P,
[1]M.Phil Scholar, Dept of computer science and Applications, KMG College of Arts & Science, Gudiyatham,
[2]Asst prof, PG and Research Dept of computer science and Applications, KMG College of Arts & Science, Gudiyatham.

Abstract:
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 big data. 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. Cloud computing plays a very vital role in protecting data, applications and the related infrastructure with the help of policies, technologies, controls, and big data tools . 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. Google has introducedMapReduce [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 and the reduce. With Hadoop it is easier for organizations to get a grip on the large volumes of data being generated each day, but at the same time can also create problems related to security, data access, monitoring, high availability and business continuity.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.

2. 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

INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 2 ISSUE 6 (2016) PAGES 1462 - 1466
RECEIVED : 24/11/2016. PUBLISHED: 08/12/2016
December 8, 2016

"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 uses networks of a large group of servers with specialized connections to distribute data processing among the servers. Instead of installing a software suite for each computer, this technology requires to install a single software in each computer that allows users to log into a Web-based service and which also hosts all the programs required by the user. There's a significant workload shift, in a cloud computing system.

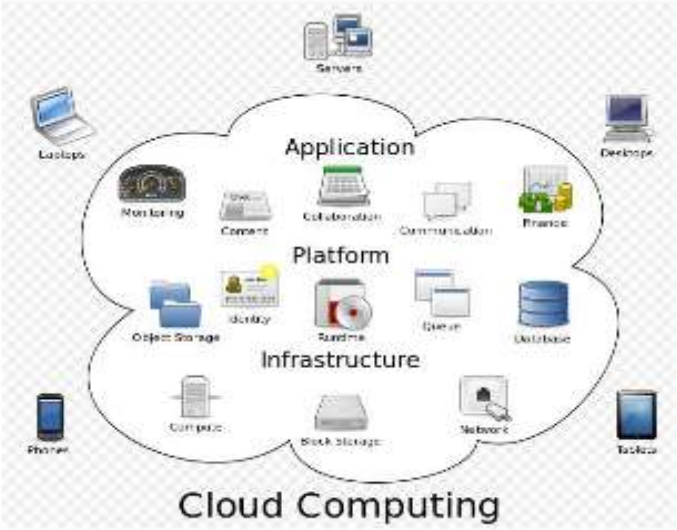

Fig.1.Cloud Computing Structure

Local computers no longer have to take the entire burden when it comes to running applications. 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'scomputer 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 real time applications which use Cloud Computing are Gmail, Google Calendar, Google Docs and Dropbox etc.,

3. RELATED WORK

Hadoop (a cloud computing framework), a Java based distributed system, is a new framework in the market. Since Hadoop is new and still being developed to add more features, there are many security issues which need to be addressed. Researchers have identified some of the issues and started working on this. Some of the notable outcomes, which is related to our domain and helped us to explore, are presented below. The World Wide Web consortium has identified the importance of SPARQL which can be used in diverse data sources. Later on, the idea of secured query was proposed in order to increase privacy in privacy/utility tradeoff. Here, Jelena, of the USC Information Science Institute, has explained that the queries can be processed according to the policy of the provider, rather than all query processing. Bertino et al published a paper on access control for XML Documents [12]. In the

paper, cryptography and digital signature technique are explained, and techniques of access control to XML data document is stressed for secured environment. Later on, he published another paper on authentic third party XML document distribution [13] which imposed another trusted layer of security to the paradigm. Kevin Hamlen and et al proposed that data can be stored in a database encrypted rather than plain text. The advantage of storing data encrypted is that even though intruder can get into the database, he or she can't get the actual data. But, the disadvantage is that encryption requires a lot of overhead. Instead of processing the plain text, most of the operation will take place in cryptographic form. Hence the approach of processing in cryptographic form added extra to security layer. IBM researchers also explained that the query processing should take place in a secured environment. Then, the use of Kerberos has been highly effective. Kerberos is nothing but a system of authentication that has been developed at MIT. Kerberos uses an encryption technology along with a trusted third party, an arbitrator, to be able to perform a secure authentication on an open network. To be more specific, Kerberos uses cryptographic tickets to avoid transmitting plain text passwords over the wire. Kerberos is based upon Needham-Schroeder protocol.

4. HADOOP DISTRIBUTED FILE SYSTEM (HDFS)

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 manufacturing. 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.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 .

5. ANALYSIS

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.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.

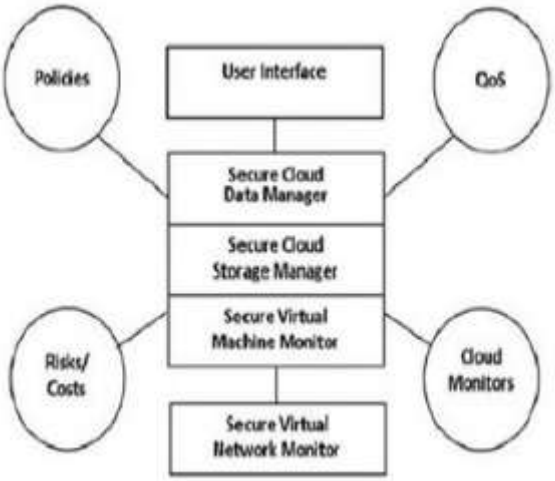

Fig.2. Layered framework for assuring cloud

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.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.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.

CONCLUSION

Cloud environment is widely used in industry and research aspects; therefore security is an important 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.