Comprehensive Security on Big Data In Cloud Environment

R.A.Tijare, M.S.Burange

Abstract— Cloud computing is the long dreamed vision of computing as a utility, where users can remotely store their data into the cloud so as to enjoy the on-demand high quality applications and services from a shared pool of configurable computing resources. By data outsourcing, users can be relieved from the burden of local data storage and maintenance. To manage the increasing need of large amount of data, big data is used. Big Data and cloud computing are two important issues in the recent years, enables computing resources to be provided services with high efficiency and effectiveness. We are focusing on these problems and 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.

Index Terms—Cloud Computing, Big Data, Hadoop, Map Reduce, HDFS (Hadoop Distributed File System).

1 INTRODUCTION

THe internet in the modern age is almost entirely driven by cloud computing model. The on-demand services and the service oriented architecture has now become essential commodity of the web. 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 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 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.

The social networking and social media have gained sizable attractions amongst the users in the current age. Almost all users, irrespective of age, profession and the locale they are from, use some or the other social networking application, very extensively. There are large numbers of social applications available for the web as well as for the mobile devices. A typical characteristic of a social application, either web or a mobile based, is having a very massive user's base. The number of users using the application may run into billions which access the application simultaneously. Another striking feature of a social networking app is the data uploaded and shared by the users. Every bit of the users' data is stored on the respective server of the application. The data uploaded and shared by the users may be in various formats like textual data, images, audio and video files. The size of the files varies as per the formats of the files uploaded. In such situation, the major problem faced application development companies is to manage the large volumes of data which is also referred to as Big Data. The best examples of such user and data intensive applications are Facebook and Whats app. Data security is yet another vital aspect to be seriously handled in the applications.

Big data can be described by the following characteristics[1]:

_ Volume
_ Variety
_ Velocity
_ Variability
_ Veracity
_ Complexity

Volume:

The quantity of data that is generated is very important in this context. It is the size of the data which determines the value and potential of the data under consideration and whether it can actually be considered as Big Data or not. The name 'Big Data' itself contains a term which is related to size and hence the characteristic.

Variety :

The next aspect of Big Data is its variety. This means that the category to which Big Data belongs to is also a very essential fact that needs to be known by the data analysts. This helps the people , who are closely analyzing the data and are associated with it, to effectively use the data to their advantage and thus upholding the importance of the Big Data.

Velocity:

The term 'velocity' in the context refers to the speed of generation of data or how fast the data is generated and processed to meet the demands and the challenges which lie ahead in the path of growth and development.

Variability:

This is a factor which can be a problem for those who analyze the data. This refers to the inconsistency which can be shown by the data at times, thus hampering the process of being able

- *R.A.Tijare is student at P.R.Pote College of Engg, Amravati, India*
- *M.S.Burange is assistant professor at P.R.Pote College of Engg, Amravati,India.*

to handle and manage the data effectively.

Veracity:

The quality of the data being captured can vary greatly. Accuracy of analysis depends on the veracity of the source data.

Complexity:

Data management can become a very complex process, especially when large volumes of data come from multiple sources. These data need to be linked, connected and correlated in order to be able to grasp the information that is supposed to be conveyed by these data. This situation, is therefore, termed as the 'complexity' of Big Data.

2 Literature Survey

Along with the increasing popularity of the Could Computing environments, the security issues introduced through adaptation of this technology are also increasing. Though Cloud Computing offers many benefits, it is vulnerable to attacks. Attackers are consistently trying to find loopholes to attack the cloud computing environment. The traditional security mechanisms which are used are reconsidered because of these cloud computing deployments. Ability to visualize, control and inspect the network links and ports is required to ensure security. Hence there is a need to invest in understanding the challenges, loop holes and components prone to attacks with respect to cloud computing, and come up with a platform and infrastructure which is less vulnerable to attacks. Big Data 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.

Hadoop, 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. *Hadoop* Map Reduce is a framework 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

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 [2]. 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 [3] which imposed another trusted layer of security to the paradigm. Moreover, 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. Airavat [4] has shown us some significant advancement security in the Map Reduce environment. In the paper, Roy and et al have used the access control mechanism along with differential privacy. They have worked upon mathematical bound potential privacy violation which prevents information leak beyond data provider's policy

3 Proposed work

In the proposed work, we aimed at providing an all round security for the users data used and shared in a social application. The security of the data would be with respect to the three main security steps Authentication, Authorization and Accounting. Besides these security aspects the application would be equipped with other necessary features like

Key encryption:

We are encrypting key and storing it in database. It will give protection from hackers.

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.

Honey pot:

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.

Data integrity:

Maintaining and assuring the accuracy and consistency of data stored in cloud environment.

Third party auditor:

Cloud computing helps in storage of data at a remote site in order 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 [1].

Data partitioning:

Data partitioning is normally done for managing data and for increasing performance or availability reasons. Each partition may be spread over multiple nodes, and users at the node can perform local transactions on the partition..

4 Conclusion

The internet has become most important medium now a day for performing almost everyday tasks. Therefore huge amount of Data Lake are being created every day. Multinational companies are using technologies like cloud computing and big data for there business. So management of information has been difficult by traditional database and providing security to information has become difficult for traditional database. With the use of big data (hadoop) analysis and management of data has become easy. Huge amount of data handling can be done with big data. By improving security of big data in cloud environment ,we can create more effectiveness in storage and management of large amount of data.

References

[1] Venkata Narasimha Inukollu, Sailaja Arsi and Srinivasa Rao Ravuri,"*Security Issues Associated With Big Data In Cloud Computing.*" International Journal of Network Security & Its Applications (IJNSA), Vol.6, No.3, May 2014.

[2] Bertino, Elisa, Silvana Castano, Elena Ferrari, and Marco Mesiti. "*Specifying and enforcing access control policies for XML document sources.*"pp139-151.

[3] E, Bertino, Carminati B, Ferrari E, Gupta A , and Thuraisingham B. "*Selective and Authentic ThirdParty Distribution of XML Documents.*"2004,pp.1263,1278.

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

[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, International Journal of Network Security & Its Applications (IJNSA), Vol.6, No.3, May 2014.

[8] 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.

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

[10] 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.