International Journal of Technology Research and Management
ISSN (Online): 2348-9006
Vol 2 Issue 9 September 2015

Securities Issues with Big Data in Cloud Computing

Dr. Vikash Kumar Singh[1], Devendra Singh Kushwaha[2], Manish Taram [3], Nikhilesh Ganvir [4]
Head (I/C), Dept. of computer Science IGNTU Amarkantak (M.P.)[1]
Assistant Professor, Faculty of Vocational Educational IGNTU Amarkantak (M.P.)[2]
Research Scholar, IGNTU Amarkantak (M.P.)[3]
Faculty of Vocational Educational IGNTU Amarkantak (M.P.)[4]
drvksingh76@gmail.com[1], devendra2904@gmail.com[2], Manishtaram86@gmail.com[3], nikhilesh.ganvir29@gmail.com[4]

Abstract: *Big data is a collection of data sets which is very large in size as well as complex. Generally size of the data is Petabyte and Exabyte. Every organization has big data. This big data contains structured, semi structured and unstructured data. Social networking users are increasing so the data of the social networking sites are also increasing rapidly. Mostly these data consists of images, videos, audios, conversations and e-mails. They are unstructured big data. So there is a need to process this data intelligent. 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

The Data volumes are increasing rapidly so processing such huge amount of data has become very difficult. Big data is defined as the five Vs that is volume, velocity, variety, value and veracity [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 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].

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 note worthy enhancements in securing the cloud that will ultimately provide us with a secure cloud.

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 Map Reduce 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].

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 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. There's a significant workload shift, in a cloud computing system. 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.

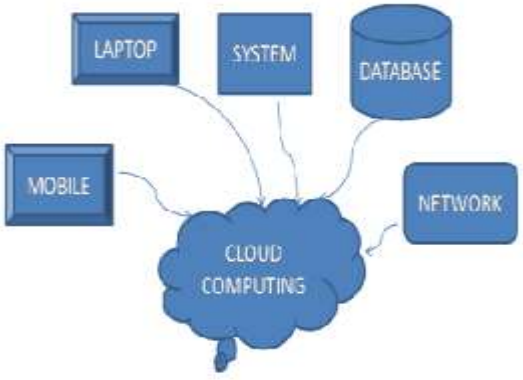

Figure 1: Cloud Network

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.

2. BIG DATA

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. The three main terms that signify Big Data have the following properties:

- Volume: Many factors contribute towards increasing Volume-storing transaction data streaming data and data collected from sensors etc.
- Variety: Today data comes in all types of formats-from traditional databases,text documents, emails, video, audio, transactions etc.,
- Velocity: This means how fast the data is being produced and how fast the data needs to be processed to meet the demand.
- The other two dimensions that need to consider with respect to Big Data are Variability and Complexity [5].
- Variability: Along with the Velocity, the data flows can be highly inconsistent with periodic peaks.
- Complexity: Complexity of the data also needs to be considered when the data is coming from multiple sources. The data must be linked, matched, cleansed and transformed into required formats before actual processing.

Technologies today not only support the collection of large amounts of utilizing such data effectively.

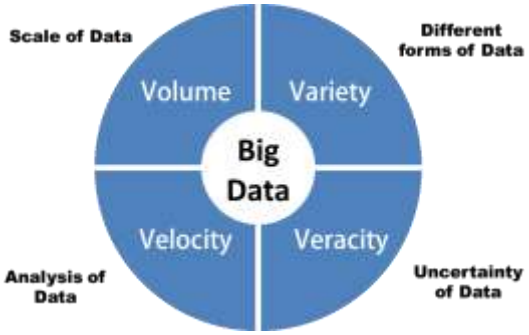

Figure 2: Properties of big data

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 the Apache project 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 of tera bytes. 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, and 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)

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.

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.

4. 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 challenges of security in cloud computing environments can be categorized into network level, user authentication level, data level, and generic issues.

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 as administrative rights for nodes, authentication of applications and nodes, and logging.

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

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

5. 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] Vikash k singh, Devendra Singh kushwaha, Shaibya Singh, Sonal Sharma “Scope of Big Data and Its Applications” International Journal of scientific research and management (IJSRM) Volume3 Issue1 2015 ISSN (e): 2321-3418, Page No. 1996-1999.

[2] Yuri Demchenko, Paola Grosso, Cees de Laat, Peter Membrey,“Addressing Big Data Issues in Scientific Data Infrastructure”,IEEE 2014

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

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