Asian Journal of Convergence in Technology *Volume1, Issue 6*
Issn No.:2350-1146, I.F-2.71

Big Data Security Issues and Challenges in Cloud Computing

Aarti A Gangawane
Department of Computer Science & Engineering
V.V.P Institute of Engineering and Technology
Solapur University, Solapur
aarrti.18@gmail.com

Anjali Devi
Department of Computer Science & Engineering
V.V.P Institute of Engineering and Technology
Solapur University, Solapur

Abstract— **In this paper we are going to discuss Big Data security issues and challenges in cloud computing, and also discuss Map Reduce and Hadoop environment used for data processing file management. In this paper we discuss Need of security in cloud computing with big data. Could Computing environments gaining popularity now a days, along with this the security issues through use of this technology are increasing. We will discuss various possible solutions for the security issues in cloud computing and Big Data. Many organizations, business, companies and many industries use the big data applications. Cloud computing security includes computer security, network security, information security, and data privacy. Cloud computing plays a very important role in protecting data, applications with the help of policies, technologies, controls, and big data tools.**

Keywords— Big Data, Cloud Computing, Hadoop, Map Reduce, HDFS (Hadoop Distributed File System), Distributed Nodes, Distributed Data, Internode Communication

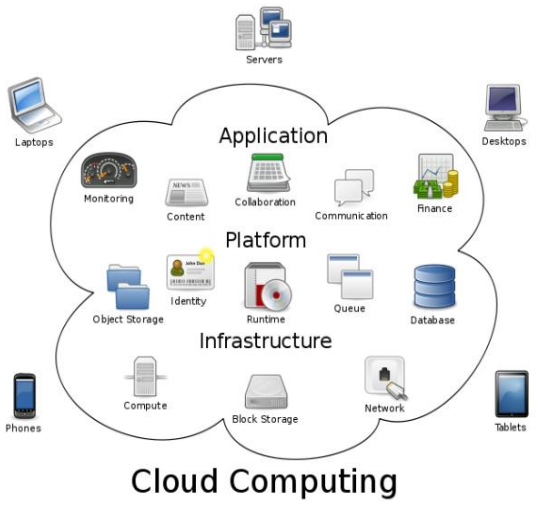
I. INTRODUCTION

Society is becoming increasingly more instrumented and as a result, organizations are producing and storing vast amounts of data. 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. Now a day's Cloud computing is becoming a very popular technology, the cloud comes with various security issues and the challenges such as owner of data may not have control of where the data is exactly placed.If one want to get the benefit of cloud computing,We need to protect the data from untrusted processes.Googles MapReduce is a parallel and distributed solution for processing large datasets on commodity hardware. Apache's HDFS (Hadoop Distributed File System) as software component is used in synchronization with cloud computing combined with MapReduce as a integrated part. Hadoop is an open source framework that allows distributed processing of large datasets including a distributed file system, provides to the application programmer the abstraction of the map and the reduce.with Hadoop it is easy to get grip on large volume of data but at the same time it creates the problem of security,data access, monitoring etc.

In this paper some approaches are mentioned that provides the security.

A. CLOUD COMPUTING

Cloud computing is recently evolved computing technology based on utility and consumption of computing resources.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 main goal of cloud computing is to increase computing power. cloud computing relies on sharing of resources over a network. 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. Cloud is classified as private, public or hybrid. A private cloud is created with business data center and provides services to internal user. In a public cloud model, a third party provider delivers the cloud services over internet. Hybrid cloud is the combination of public cloud and private cloud.



B. BIG DATA

Big data is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications. The challenges include capture, curation, storage, search, sharing, transfer, analysis, and visualization. Big Data have the following properties.

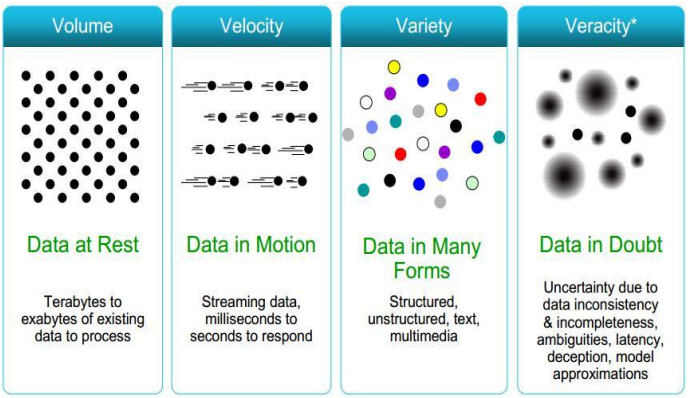


Fig: Properties Of Big Data.

Variety : Variety represents the data types i.e. traditional databases, text documents, emails, video, audio, transactions etc.
Velocity: It refers to the rate at which the data is produced and processed.
Volume: volume defines the amount of data.
Veracity: Veracity refers to how much the data can be trusted given the reliability of its source.
Value: value corresponds the monetary worth that a company can derive from employing Big Data computing.

C. HADOOP

Software platform that lets one easily write and run applications that process vast amounts of data. It is a part of the Apache project sponsored by the Apache Software Foundation. Here's what makes it especially useful:
Scalable: It can reliably store and process petabytes.
Economical: It distributes the data and processing across clusters of commonly available computers (in thousands).
Efficient: By distributing the data, it can process it in parallel on the nodes where the data is located.
Reliable: It automatically maintains multiple copies of data and automatically redeploys computing tasks based on failures.

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).

D. MAP REDUCE

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. Sort/merge based distributed computing. Initially, it was intended for their internal search/indexing application, but now used extensively by more organizations (e.g., Yahoo, Amazon.com, IBM, etc.)It is functional style programming (e.g., LISP) that is naturally parallelizable across a large cluster of workstations or PCS.
The underlying system takes care of the partitioning of the input data, scheduling the program's execution across several machines, handling machine failures, and managing required inter-machine communication. (This is the key for Hadoop's success)

E. HADOOP DISTRIBUTED FILE SYSTEM

HDFS 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.

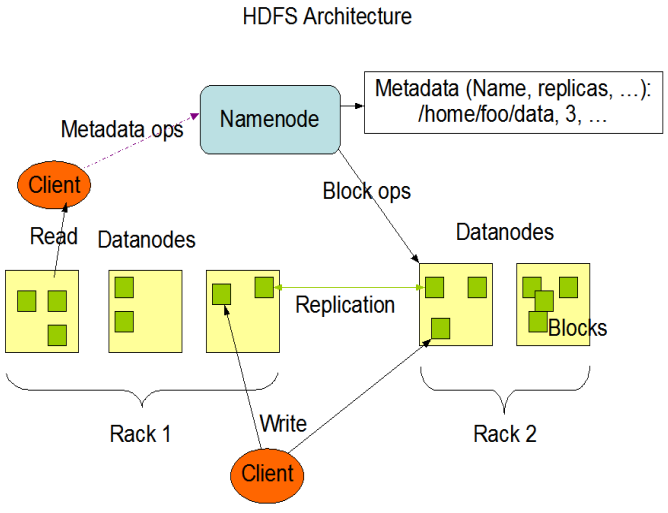


Fig. HDFS Architecture.

F. NEED OF SECURITY IN BIG DATA

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 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 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, and 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.

II. 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 are 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. 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 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 forms 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 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. The above works have influenced us, and we are analyzing various approaches to make the cloud environment more secure for data transfer and computation.

III. ISSUES AND CHALLENGES

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, Internodes 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.

A. DISTRIBUTED NODES

Distributed nodes 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 can happen 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.

B. DISTRIBUTED DATA

In order to alleviate parallel computation, a large data set can be stored in many pieces across many machines. Also, redundant copies of data are made to ensure data reliability. In case a particular 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 data are copied to another node/machines based on availability and maintenance operations. In traditional centralized data security system, critical data is wrapped around various security tools. This cannot be applied to cloud environments since all related data are not presented in one place and it changes.

C. INTERNODE COMMUNICATION

Much Hadoop distributions use RPC over TCP/IP for user data/operational data transfer between nodes. This happens over a network, distributed around globe consisting of wireless and wired networks. Therefore, anyone can tap and modify the inter node communication for breaking into systems.

D. DATA PROTECTION

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

E. ADMINISTRATIVE RIGHTS FOR NODES

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

F. AUTHENTICATION OF APPLICATIONS AND NODES

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

G. LOGGING

In the absence of logging in a cloud environment, no activity is recorded which modify or delete user data. No information is stored like which nodes have joined cluster, which Map Reduce jobs have run, what changes are made because of these jobs. In the absence of these logs, it is very difficult to find if someone has breached the cluster if any, malicious altering of data is done which needs to be reverted. Also, in the absence of logs, internal users can do malicious data manipulations without getting caught.

IV. THE APPROCHES FOR SECURITY

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. 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.

A. 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.

B. 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.

C. 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.

D. SOFTWARE FORMAT AND NODE MAINTENANCE

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

E. 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.

F. 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.

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

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

[3] 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.

[4] J. Dean and S. Ghemawat, ``MapReduce: Simplied Data Processing on Large Clusters,'' OSDI 2004. (Google)

[5] F. Schomm, F. Stahl, G. Vossen, Marketplaces for Data: An Initial Survey, SIGMOD Record 42 (1) (2013) 15-26.

[6] The Intel science and technology center for big data, http://istc-bigdata.org

[7] McAfee, E. Brynjolfsson, Big Data: The Management Revolution, Harvard Business Review (2012) 60-68

[8] The Age of Big Data. Steve Lohr. New York Times, Feb11, 2012. http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html