Security Issues Associated with Big Data

Mohammed Asrar Naveed
Department of ISE
Acharya Institute of Technology
Bengaluru, India

Chaitra B
Department of ISE
Acharya Institute of Technology
Bengaluru, India

***Abstract*- Big data is well-defined as huge volume of data which involves new skills so that it becomes likely to extract value from it by analysis process. Due to large data size, it becomes very hard to make effective study. In this paper the main focus is on security issues with big data in cloud computing and various possible solutions for the issues are also discussed. Security in cloud computing involves processor security, grid security, information security and data privacy. As cloud computing technology is being cast-off to decrease the usage price of computing resources, many establishments grew interest of travelling their old organization to the cloud computing system and since big data is a recent forthcoming technology in the market used for analytics purpose it would bring massive paybacks to the organization.**

Keywords: - Hadoop, Mapreduce, HDFS.

I. INTRODUCTION

Data is growing at a huge rate making it difficult to handle such a large amount of data .The main difficulty in handling such large amount of data is mainly because the volume is increasing rapidly in comparison to the computing resources. In order to analyze huge and compound data and to recognize patterns, it is necessary to securely store and manage huge amount of complex data where in the Big data term comes into picture. There are different definitions for the term big data. The most popular definition is given by their four characteristics called "4V": volume (the data volumes are very huge which cannot be processed by normal methods) Velocity (The data is generated with the greater velocity and must be stored and processed rapidly) Variety (Variety of Data types: structured, Semi structured and Unstructured) Variability (along with velocity, the data flow can be highly inconsistent).

Cloud comes with an obvious security task i.e the vendor of the data might not have any control of where the data is stored. Henceforth it is compulsory to guard the data within the unreliable process.

Google has introduced MapReduce framework for analyzing huge amounts of data. Apache Hadoop distributed file system (HDFS) is a software element for cloud computing which also includes MapReduce. With Hadoop it is at ease for administrations to get a grip on huge sizes of data being produced each day but at the similar period it can also make hitches associated to security.

In this paper, we come up with some methods in providing security. We ought schemes that can be clever to route huge and immense volumes of data. Hadoop framework merges k-mean technique and data mining technology to resolve some of the security issues.

A .Hadoop

Hadoop is an open source venture presented by Apache software which consists of HDFS, MapReduce, Hive, Zookeeper and other ventures.

Hadoop mostly comprises of

- File system (The Hadoop file system)
- Programming standard (Map Reduce)

The extra substitute projects provide complementary service as Hadoop also lacks safety measures. Kerberos was integrated into Hadoop in 2009 by Yahoo. The user has to obtain access certification as Hadoop cluster is a Master/slave erection. Using Hadoop huge data groups can be processed across a collection of servers and application can be made to run on the structures with thousands of lumps involving thousands of terabytes. Hadoop is used by Google, Yahoo and Amazon etc.

B.Map Reduce

MapReduce is the software design paradigm [6] allowing huge scalability. The MapReduce mostly do two unlike tasks i.e Map Task and Reduce Task.

Map tasks are specified input from distributed file system. The map task yields an order of key-value duos from the input. When a job is submitted to the MapReduce framework, MapReduce will divide it into several Map tasks and assign them to different nodes for running. Every Map task deals only with a part of the input data. After Map task processing, those intermediate state key-value pairs will be sent to the Reduce function. Reduce function will merge the pairs based on a specific key [9], then generate the output value-keys that client requires.

C. Hadoop Distributed File System (HDFS)

HDFS [7] is a file system that extents all the nodes in a Hadoop band. HDFS progresses gradually by copying data across numerous sources to overcome node failures.

D. Big data applications

The Big data application spreads across huge data groups. Data examination bowed into a difficult problem in many areas of Big data. Major features of Big data are Google's map reduce framework and apache Hadoop in which these features generates a large amount of data. Two major big data applications are Manufacturing and Bioinformatics. The infrastructure for transparency in manufacturing industry is provided by the Big data. The sensory data and the historical data constructs manufacturing sector. Hadoop has one more application in bioinformatics which covers the biological domains and the next generation sequencing.

E. Big data advantages

The software packages and the tools provide options where an individual could map the whole data across the company. This is one of the big advantages in the Big data. Big data provides a platform such that it can be customized based on user interests which can be further analyzed with analytical tools, data analytics is one of the emerging advantage in the Big data.

II. MOTIVATION AND RELATED WORK

A. Motivation

As with the growing admiration of the Could Computing, the security issues along with that technique is also increasing. Though this cloud computing offers many benefits, it is very sensitive to attacks. Attackers are reliably trying to find loop holes to attack the cloud computing atmosphere. The outdated security mechanisms are reviewed because of these cloud computing arrangements. Capability to imagine, control and examine the network links and ports is vital to confirm security. Hence there is a need to realize the loopholes, the security threats in the cloud computing.

B. Related Work

Hadoop is a cloud computing framework and is based on java distributed system. It is a new emerging framework in the market. There are many security issues that need to be addressed in this technique and are listed below.

The Knowledge of secured query was projected in order to raise the privacy. Jelena explained that the queries can be processed based on the policy of the provider, instead of all the query.

Bertino et al proposed an access control for XML Document [2]. In this the author had used the cryptography and digital signatures for the security of data. In authentic third party system, XML document distribution [3] adds another layer of security for the data.

Kevin Hamlen and et al suggested that the data in the database should be stored in encrypted form rather than the plaintext form.

The advantage of this is even if any malicious node extracts the data it is difficult to get the useful information.

The drawback is that encryption would impose a lot of overload.

Roy and Airavat [4] have used the access control mechanism along with differential secrecy for data protection and to prevent the leak of the data, they worked on the mathematical bound potential privacy violation.

III. ISSUES AND CHALLENGES

The challenges of security in Big data environments can be categorized into four; Network level, User authentication level, Data level, and Generic issues [1].

Network level: The challenges that deal with network protocols and network security are distributed nodes, distributed data, Internode communication.

Authentication level: There are several issues related to encryption/decryption techniques at the user level and tracking log files at the administrative level.

Data level: There are issues related to data integrity and availability in which data protection is a major challenge along with monitoring the distributed data.

Generic types: Deals with the challenges associated with primitive security tools.

A.Distributed Nodes

It is an architectural issue where in the data analysis can be done in any of the cluster hence it is hard to find the particular scene of computation. It is very difficult to provide the security for the data where the computation is occurring.

B. Distributed Data

In this the large data is partitioned into small pieces across many machines and the duplicate copies are also made. In case any copy is corrupted then the redundant copy can be retrieved. In a huge computing environment it is difficult to locate the copy of a particular file. To overcome this, related technologies are used.

C. Internode Communication

To transfer the user data between the nodes the Hadoop distribution uses RPC over TCP/IP. This happen over a network: Wired or Wireless, therefore any one can enter into the network and make changes in the internode communication.

D. Data Protection

Data stored in the form of plain text i.e without encryption to increase the efficiency in Hadoop, if any malicious node attacks the data then there is no way to stop the malicious node.

E. Administrative Rights for Nodes

A node can access the data if it has the administrative rights [8].The uncontrolled access leads to damage the data and any other malicious node can attack the data.

F. Authentication of Applications and Nodes

Many nodes can combine to form a cluster, if authentication is not done at the node level then there are chances of hackers entering into the clusters.

G. Logging

The activities which are not recorded in the system does not provide any information regarding, which node have joined the cluster and what changes have been made in the data and even it is not possible to track the jobs done by the map reduce.

H. Traditional Security Tools

Traditional security tools are used where scalability is not huge and they cannot be directly applied to the distributed form

I. Use of Different Technologies

In cloud computing, many technologies are used. Due to huge number of technologies, minor security weakness in one node or component can bring down the whole component because of the Diversity.

IV. THE PROPOSED APPROACHES

In this session the various security measures are introduced to improve the security of the data. The proposed solutions will encourage the use of multiple technologies. Security recommendations are designed in such a way that they do not drop the efficacy and scalability.

A. File Encryption

Since the data is stored in the nodes in a cluster, an inducer or a malicious user can get the user data. Therefore the data stored in the node is to be encrypted in order to protect from the hackers. Different Encryption keys are to be used in the Encryption algorithm and the keys are stored securely, henceforth even if the hacker accesses the node and get the user data, he cannot extract the useful information.

B. Network Encryption

The communication over network should be encrypted. SSL is to be considered for RPC procedure calls, so that even if a malicious node accesses network communication packets, it cannot extract useful information or alter the packets.

.

C. Logging

The jobs occurred in the system should be logged and also the information used in the jobs are also logged and these logs are to be viewed regularly in order to check whether any manipulation is done or any malicious node has entered into the system.

D. Software Format and Node Maintenance

Software which runs on the node should be regularly formatted in order to eliminate the virus which is present and Hadoop software must be updated eventually to increase the performance and make the system more secure from the hackers.

E. Nodes Authentication

In order to restrict the malicious node to enter and manipulate data, the node authentication must be done. For such events many techniques such as Kerberos must be used.

F. Rigorous System Testing of Map Reduce Jobs

When a designer carves a map reduce job, it is meticulously tested in a distributed atmosphere instead of testing in a single machine in order to get the Robustness and stability.

G. Honeypot Nodes

Honeypot node is like a regular node which should be present in the cluster of nodes, the honeypot nodes determines the malicious node in the cluster and eliminates it by performing some necessary actions.

H. Access Control

The security policy of the sensitive data is controlled by the Data providers by using JVM and Map Reduce framework enforcement of privacy. It has inbuilt applications which stores the pool of identities.

Real time access control is used in addition to the access control mechanism. By assigning label to the data, the data can be protected through the label security method. Further if the user label is matched with the data label then the access is permitted.

I. *Third Party Secure Data Publication to Cloud*

In order to maximize the resource utilization cloud computing permits to store the data in a remote places. Therefore, it is very necessary for the data to be protected and only authorized individuals should have access to that data. Henceforth for data outsourcing and for external publications the hold of authority should be given to the third party [5]. The machine serves as third party publisher in the cloud environment which holds the sensitive data. This data needs to be protected and the beyond conversed techniques have to be utilized to ensure the maintenance of authenticity.

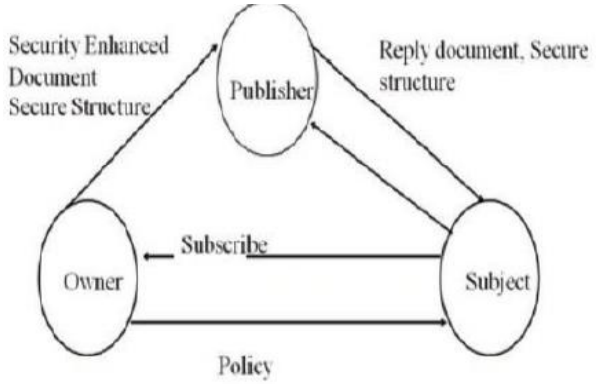

Fig.1: Third party sheltered data publication realistic to cloud.

V. CONCLUSION

Cloud atmosphere is broadly used in commerce and research facets; therefore security is a vital facet for organizations seriatim on these cloud environments. Using proposed methods, cloud atmospheres can be secured for intricate business operations.

REFERENCES

[1] 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.

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

[3] E, Bertino, Carminati B, Ferrari E, Gupta A , and Thuraisingham B. "Selective and Authentic Third-Party 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] 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.

[6] 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.

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

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

[9] Jeffrey Shafer, Scott Rixner, and Alan L. Cox. The Hadoop Distributed Filesystem Balancing Portability and Performance[R]. IEEE, 2010.

.