Journal of Engineering Technology

Vol. 6, Special Issue (Emerging Trends in Engineering Technology) Mar. 2018, PP. 132-138

RESEARCH ARTICLE -ENGINEERNG TECHNOLOGY

Recent advances in big data storage and security schemas of HDFS: a survey

Balaraju.J[1] and P.V.R.D Prasada Rao[2]

[1]Research Scholar, School of Computer Science & Engineering,
Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh, India – 522502
[2]Professor, School of Computer Science & Engineering, Koneru Lakshmaiah Education Foundation,
Vaddeswaram, Guntur, Andhra Pradesh, India – 522502

Abstract: Technological revolution in the digital world has led to huge data generation from many sources in various formats, which results in changing its native data type (Homogeneous to Heterogeneous). The digital data has increased approximately 350 times from 140 Exabyte's in last decade to 50,000 Exabyte's in ending of 2020. As an experiment of the technological improvement, the increasing big data becomes major issue in the government sector, science, research and business enterprises. The data storage, processing and security is difficult by using the conventional database tools and database management systems at traditional cloud centers. To overcome all these challenges, the management of huge data is to be done by maintaining Hadoop based cloud data centers. This paper has two sections, section-1 explains different big data storage, security mechanism and HDFS security mechanisms. Section-2 finds out recent advances in Hadoop for securing the big data in cloud. This paper ultimately examines and studies storage, security and privacy mechanisms and suggests a solution for securing huge data in cloud by using Hadoop eco systems.
Keywords: Big Data, HDFS Security, Hadoop, Data Storage, Security Mechanisms.

1. Introduction and Characteristics of Big Data

Data which is beyond the storage and processing capacity is called as big data [1]. The various issues of big data can be explained by using 3V's which are typically meant to characterize the big data as Volume, Velocity, and Variety. Volume refers to quantity of data, Velocity refers to the speed of generating data and Variety is about consisting of different categories of data such as structured, semi-structured and unstructured. The above big data characteristics are helpful for extracting its hidden patterns. These characteristics make it difficult for storing and processing massive data using Traditional information processing applications. Hadoop is one of the powerful tool for handling big data and it stores the data in a distributed manner and also performs parallel processing of data. Security is a tradeoff in hadoop since it does not have its own security protocol rather it uses external security mechanisms like Kerberos for authentication and Bull eye for securing sensitive data in hadoop clusters.

2. Hadoop

Hadoop [2] is meant for storing and processing big data in a computing environment which is distributed in nature. Hadoop is parallel and distributed File System for the purpose of storing huge data which groups with constellation of low cost hardware and with streaming access configuration. Hadoop follows a principle called write once and read many times, but the content of the file remains unchanged. Hadoop undertakes mainly two tasks they are, storing of huge data in Hadoop Distributed File System [3][4] (HDFS) and parallel processing called Map Reduce[5]. Hadoop stores the data as it is without encryption to mend proficiency. HDFS supports running access to file system data and it affords substantiation, seclusion and security.

2. 1. Storage, Security of Big Data using HDFS.

HDFS was established using distributed file system design that provides high-performance to data across hadoop clusters. Unlike other distributed systems, HDFS has high fault tolerance and also supports any type of hardware in its cluster ranging from low to high cost. HDFS comprises comprehensive data and provides access easily. To store such enormous data, the files are stowed through numerous systems. HDFS requires applications for parallel processing such as MapReduce. The input file is first split into blocks of equal size but the last block which is then replicated across Data Nodes. Currently, default block size is 128MB to 512MB which was previously 64MB and default replication factor is 3. Blocksize and replication factors are configurable parameters. Hadoop distributed procedure framework is not developed for security concern. Later, Hadoop has become a widespread platform, as an outcome security precautions have started to advance. Hadoop chains trusted environment between client (user) and Name Node which is module of HDFS and copes the Data Node. Many arbitrary encryption methods are smeared on data which is deposited in Data Node for securing data. The user himself need to authenticate Name Node in order to access.

3. Review of related works

In the literature survey, numerous approaches have been anticipated for the Storage, Refuge and Privacy Stabilizing of big data in Cloud, and security perception by using Hadoop. Among the most lately published works are those presented as follows. This is divided into two categories i.e. storage techniques, Security and Secrecy methods.

3.1. Big data Storage Perspective:

Jian Jun Luo a, LingyanFan et al.[6] have explained hardware based **S**elf-search **S**torage **D**evice beside with search engine which is entrenched with big data stowage devices like RAID and SAS. FPGA (Field Programmable Gate Array) was firstly applied to arrange the SSD disk with fixed iSearch engines but its scheduling parameter produced the congestion for high speed throughput. An assessment was done by smearing 24 iSearch fixed SSD units and the recital was close to 1GB/s corresponding to the supreme speed of the iSCSI interface associated by 10 GB/s Ethernet with server. A disk array was fabricated with 24 SSD units and each unit had 1TB density. A 24 TB database was encumbered to run the assessment program. An entire inquiry without iSearch motor is included around 24s, and hunt with iSearch motor is at a normal of 5.5ms. The more target substance were scattered in database by utilizing iSearch motor and extremely predominant spread in database.

Hanadi Ahmed Hakami, Zenon Chaczko and Anup Kale et al [7] have explicated the data storage in DNA categorization prospects and the DNA stockpiling is appeared to be exceptionally viable. This research in DNA storage recommends that it can sanction for far more seclusion and security by paralleling with silicon devices. The coding method is distinguished and used to store information and examined by Cell, Risca and Bancroft. There are various encoding measures to store information under DNA movements by utilizing DNA codes. An interested Haffman capable source coding strategy is adept to decipher. DNA lattice has been utilized to epitomize the Metadata and it follows by renovating the DNA lattice into Quick Response (QR) portrayal that arrangements an expansive extent of reasonable standard. Subsequently this plan is profoundly relevant for taking care of and stowing of monstrous measures of endless sorts of information. This type of storage system is in premature platforms of research.

Journal of Engineering Technology

Vol. 6, Special Issue (Emerging Trends in Engineering Technology) Mar. 2018, PP. 132-138

RESEARCH ARTICLE -ENGINEERNG TECHNOLOGY

3.2. Security and Privacy Perspective

In Pradeep Adluru and Srikari Sindhoori Datla et al. [8] have illustrated the safety and protection of Bigdata in Hadoop eco frameworks. In this method The User himself authenticate to the Name Node in order to access data. A trustee mechanism is shaped between user and Name Node. The Name Node and Data Node have the in place control over information yet client does not have control over information. Only Authenticated user can access data directly, others are not permitted. The safe figures or encryption strategies that are actualized in this framework are RSA, Rijndael, AES and RC6. The encryption and decoding of the data are taken care of in the Map Reduce module. The SHA-256(Secured Hash Algorithm) hashing technique is used for authentication between users of Name Node by providing hash function by Name Node.

In, B. Saraladevia, N. Pazhanirajaa, P. Victer Paula et al[9] have explicated security disputes in HDFS. The HDFS is the base layer of Hadoop Architecture which contains disparate game plans of information and it is more touchy to security concerns. The following three security and privacy appliances are based on HDFS. We can progress security in big data by using any one of the following approaches or by combining these three approaches in HDFS.

a). In order to accomplish the authentication between user and Name Node for accessing data blocks from Data Node, Kerberos authentication mechanism is used for improving the security. The association between client and namenode is reached by using remote procedure call[10] in HDFS. User want to access data from Data Node, they must use Ticket Granting Ticket or Service Ticket to authenticate by Name Node. The TGT and ST can be renewed while Kerberos is rehabilitated after long running of jobs, new TGT and ST also issued and strewn to all tasks. Key Distribution Center (KDC) issues the Kerberos Service Ticket using TGT after getting request from task. The aim of this ways and means is to create endorsement between user and Name Node for retrieving data.

b). In order to screen and secure profound data like credit card numbers, passwords, account numbers, personal details and to upturn the haven in Hadoop base stratum, the Bull's Eye Approach is proposed. This exemplary outlooks all the sensitive info in 360^0 and find whether all warehoused data is secured without any menace. It consents only accredited users to realm the personal information in a right way. Bull's eye approach increases security in Hadoop base stratum used in HDFS for providing security in 360^0 from node to node. By applying this tactic in Data node of rack1, and it checks whether the data is put in storage properly in data block without any peril, and allows only the particular punter to store in required blocks. This is also bridge gap between original data and replicated data. When the patron wants to rescue data from replicating data nodes it is also preserved by "Bull Eye Approach" and it checks whether there is a proper relation between two data racks. This can be instigated below the data node where the patron read or writes the data to store in blocks. It is not only employed in the rack 1 similarly it is implemented in Rack 2 in order to upturn the security of the lumps inside the data nodes in 360°. This algorithm travels from less terabytes to multi petabytes of semi structured structured and unstructured data and stored in HDFS layer in all angles. Many companies like Data guise's DGsecure and Amazon Elastic MapReduce used this approaches. DG secure Company gave a Data driven and Governance arrangements and furthermore includes security for Hadoop in the cloud. Information appearance Company chose to keep up and give security in hadoop any place it is situated in cloud.

c). In HDFS, if the Name Node is lost entire Hadoop system will fail. The Name Node is Master Node which vittles metadata and user cannot admittance their data without Name Node. With a specific end goal to upsurge security in information attainable quality is achieved by including Secondary Name Node. Two Name Node servers are supported to run beneficially in a similar group and two excess Name Nodes are conveyed by Name Node Security Enhance (NNSE), and it holds bull eye algorithm. Hadoop Administrator can access

both nodes, one node act as master node and other act as slave node. The master Name Node in order to clatter the administrator, it needs to ask NNSE to prove data from slave node in order to recuperate unavailability of data in fortified manner. Admin cannot access data from slave node without approval from NNSE and it may condense complex recovery issue. If Name node acts as a master node, there transpires a incessant risk, diminishes secured information accessibility and bottleneck execution over a neighborhood or Wide Area Network. Along these lines in future we can likewise rise withdraw by utilizing imperative setup that gives and guarantees information accessible to customer in a secured route by duplicating many Name hub by Name Node Security Enhance in HDFS hinders between voluminous server farms and groups.

Recent Advances Big Data Security and Storage in HDFS.

Author	Proposed Schema	Advantages
Prashant Johri et al[11]	Hadoop Security framework for perceptive and high-speed data storage and accessibility on big data platform. Encryption is performed on data using Map Reduce.	Provides additional security constraints such as Dynamic One Time Code and policies that bind with data and flexible to be used on any application. By using Map Reduce for encryption it provides faster encryption and decryption time.
Michael Kenyeba et al[12]	Discussed issues related to Hadoop Security in cloud environment and discussed possible solutions to those issues in hadoop security and authentication.	No External authentication mechanism is required for small clusters but it is not applicable to large clusters.
Bikash Agrawal et al[13]	Modification of NameNode using CheckerNode for secure deletion.	By implementing checker node with addition to the name node, checker node eradicates the problem of maintaining false copies of data at different data nodes which are not in the name node. Deletion is taken care by checker node at regular frequent intervals of time. Checker node tracks undeleted data blocks and deletes it.
Bowen Tian et al[14]	Allocation of Data depending on its type in order to improve security.	Performance levels are always little less in heterogeneous databases when compared to homogeneous database. So, to scale up performance following a data allocation scheme would result in improving both performance and security levels such as SecHDFS. SecHDFS allocates fragments of a file to different types of DataNodes, there by resolving the major risk issue caused by data replication in HDFS.
Hua Xu et al[15]	Proposed new method for location aware data block and their allocation by solving disadvantages in virtual nodes of HDFS based cloud.	The proposed schema improves performance and data reliability for HDFS based application in cloud by eliminating the co-location of virtual machine.

Yoon-Su Jeong et al[16]	A token-based authentication scheme that protects sensitive data stored in HDFS against replay and impersonation attacks.	Provides more protection to sensitive HDFS data without overburdening authentication operations.

The analysis of above table is that hadoop is mainly facing authentication and metadata security problems. The authors have tried to improve the above security features but each author focused only on one area leaving the rest behind which is a bottleneck for hadoop security.

4. Recommendation to Integrate HDFS cluster with Secure Node.

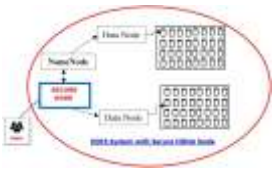

Fig.1 : HDFS Integrated with Secure-Node

Figure 1 Shows HDFS integrated with Secure-Node which is used for both authentication and meta data security purpose which leads to less computational time and security performance is enhanced. In existing system two nodes are required, one for authentication and another for metadata security. The proposed system contains only one node both for authentication and metadata security.

Table 1: Nodes and Computational time

	Existing	**Proposed**
Secure Nodes required	2	1
User computational time	2	1

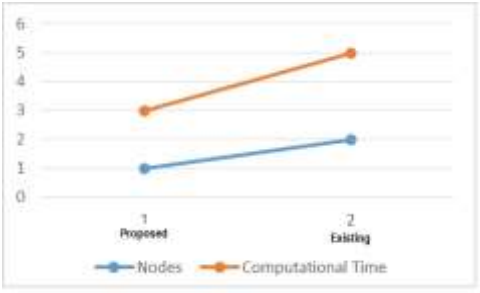

Fig 2. Representation of Nodes and Computational Time.

5. Conclusion and future work

The Storage, security and protection for Big data administrations is exceptionally thought - revoking. The solution to the above can be provided by Hadoop framework rather than any other big data tool. There is a rapid increase in the security features of Hadoop from version to version since last decade. According to section-2 of this paper Hadoop is still facing Data leakage, lack of metadata security and authentication issues. To overcome all the above problems, a complex security mechanism like DNA cryptography can be used to give optimal solution. Converting the entire big data into DNA for storage and security purpose is highly computational task. In this particular area, the research is still going on and further there may be scope in our research. One can try for the optimal solution for this problem. Instead of converting entire data into DNA form, it's better to develop DNA based authentication for accessing HDFS which can give optimal solution for fortifying data in HDFS frame work. This may jettison NNSE hindrances for security metadata in name node of HDFS.

References

[1]. YojnaArora, Dinesh Goyal, "Big Data: A Review of Analytics Methods & Techniques", International Conference on Emerging Trends in Computing and Communication Technologies (ICETCCT), 978-1-5090-5256-1116/$31.00 ©2016 IEEE.

[2]. Dhole Poonam B, GunjalBaisa L, "Survey Paper on Traditional Hadoop and Pipelined Map Reduce", International Journal of Computational Engineering Research Vol 03, Issue12.

[3].Dean, J., Ghemawat, S.: MapReduce: simplified data processing on large clusters. In: Proceedings of Sixth Symposium on Operating System Design and Implementation (OSDI04), pp. 137–150 (2004)

[4].Shvachko, K., Kuang, H., Radia, S., Chansler, R.: The Hadoop distributed file system. In: Proceedings of the 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST), pp. 1–10 (2010)

[5]. J. Dean and S. Ghemawat, Mapreduce: Simplified data processing on large clusters, Commun. of ACM, vol. 51, no. 1, pp. 107-113, 2008.

[6].Jian Jun Luo a, LingyanFan a, n, ZhenhuaLi b, ChrisTsu c "A new big data storage Architecture with intrinsic search engines" 0925-2312- 2015 Elsevier, Neurocomputing181 (2016) 147–152.

[7]. Hanadi Ahmed Hakami, Zenon Chaczko and Anup Kale "Review of Big Data Storage Based on DNA Computing", of 978-1-4799-7588-4/15 $31.00 © 2015 IEEE.

[8]. Pradeep Adluru and Srikari Sindhoori Datla "Hadoop Eco System for Big Data Security And Privacy" 978-1-4577-1343-9/12/$26.00 ©2015 IEEE.

[9].B. Saraladevia, N. Pazhanirajaa, P. Victer Paula, M.S. Saleem Bashab, P. Dhavachelvanc "Big Data and Hadoop-A Study in Security Perspective" 2015 the Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license.

[10].Heindel L.E, "Highly reliable synchronous and asynchronous remote procedure calls", Conference Proceedings of the IEEE Fifteenth Annual International Phoenix Conference on computers and communications,1996.

[11]. PrashatJohri , Arun Kumar, Sanjoy Das "Security framework using Hadoop for Big Data" , ISBN:978-1-5090-6471-7/17/$31.00 ©2017 IEEE268.

[12]. Michael Kanyeba and Lasheng Yu , "Securing Authentication Within Hadoop" , © 2016. The authors - Published by Atlantis Press, ICEMIE 2016.

[13]. Bikash Agrawal⇤, Raymond Hansen, Chunming Rong, Tomasz Wiktorski, "SD-HDFS: Secure Deletion in Hadoop Distributed File System", 10.1109/BigDataCongress.2016.30, **IEEE Xplore:** 2016.

[14]. Bowen Tian, Yun Tian, Yijie Sun, Trevor Hurt, Brandon Huebert, Waymon Ho, Yuting Zhang, Danqi Chen "A Secure Data Allocation Solution for Heterogeneous Hadoop Systems: SecHDFS", 978-1-5090-5252-3/16/$31.00 ©2016 IEEE.

[15]. Hua XU , Weiqing , gaunsheng Shu and Jing Li " Location-aware Data Block Allocation Strategy for HDFS-base Applications in the Cloud" , 2159-6190/16 $31.00@ 2016 IEEE DOI 10.1109/CLOUD.2016.40.i.

[16]. Yoon-Su Jeong · Yong-Tae Kim "A token-based authentication security scheme for Hadoop distributed file system using elliptic curve cryptography" , J Comput Virol Hack Tech (2015) 11:137–142 DOI 10.1007/s11416-014-0236-5.