IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Vol.-6, Issue-11, Nov 2018 E-ISSN: 2347-2693

Security and Privacy Issues in Big-data Hadoop: A Review

Sathisha M S[1]*, K C Ravishankar[2]

[1]Dept. of Computer Science and Engineering, Canara Engineering College, Mangaluru, Karnataka, India
[2]Dept. of Computer Science and Engineering, Govt. Engineering College, Hassan, Karnataka, India

*Corresponding Author: sathishams1983@gmail.com

Available online at: www.ijcseonline.org

Accepted: 25/Nov/2018, Published: 30/Nov/2018

***Abstract*—** Nowadays Data is one of the important assets for industries in almost all fields. Advances in information technology and its widespread growth in several areas such as business, engineering, medical, and scientific studies are resulting in data explosion. This data explosion created a new problem that, data cannot be handled by traditional techniques. This problem was, therefore, solved through the creation of a new paradigm: Big Data. Big-data refers to massive volume of structured, semi structured and unstructured data that conventional data management methods are incapable of handling. One of the finest and most popular technology available for handling and processing that enormous amount of data is the Hadoop ecosystem[4]. Enterprises are increasingly relying on Hadoop for storing their valuable data and processing it. However, Hadoop is still evolving. Many researchers have found vulnerabilities in Hadoop, which can question the security of the sensitive information that enterprises are storing on it. In order to prevent these concerns from becoming real harms, effective policy and technological measures are required on the part of organizations that uses "Big Data", as well as for individuals to whom the data relates. It is almost impossible to carry out detailed research into the entire topic of security. In this paper we try to present a big picture of the main problems related to security in a Hadoop ecosystem, along with the possible solutions to those problems proposed by the research community.

Keyword- Big data; Hadoop; security; privacy;

I. INTRODUCTION

In recent days, the word Big Data is used almost everywhere. The term Big data was introduced by Roger Magoulas from O'Reilly media in 2005[23], refers to a huge variety of big facts units almost not possible to control and manner the use of conventional information control tools – because of their length, and their complexity. The total amount of data in the world was 4.4 zettabytes in 2013. That is set to upward thrust steeply to 44 zettabytes through 2020. This sort of storage allows industries to extricate required knowledge on the right time and even in real time. Not handiest can big groups find the money for Big Data, however small organizations also can reap blessings from using this Big Data However, processing and analyzing such massive quantity of heterogeneous records is not possible through the use of established databases and traditional methods, because it requires big parallel processing strategies, as a consequence, it turned into mandatory to introduce novel strategies and tools, which might be properly aligned with this evolution. One of the best and most famous technologies to be had for dealing with and processing that huge quantity of data with the Hadoop environment. Enterprises are more and more on Hadoop for storing their reliable information and process it. Although, Hadoop continues as emerging field. There is a large vulnerability found in Hadoop[4], which might make to think of the safety of the sensitive data that enterprise area unit storing thereon

Fig 1: Big data Architecture **[1]**

The fashion in Big Data in the direction of multiplied collection, storage, and linkage of personal facts units raises protection concerns relating to theft, institutional misuse, and re-identity, in addition to unauthorized get admission to, redistribution, and repurposing of personal facts. In order to rescue from these issues from turning into real harms, powerful policy and technological measures are required on the part of organizations which are using "Big Data", as well as for individuals to whom the statistics relates. With respect to the Working Group of Big Data on the alliance on Cloud Security Company there are four exclusive elements of Big Data protection: Data privacy, infrastructure security, Data Management, and Integrity and Reactive Security.

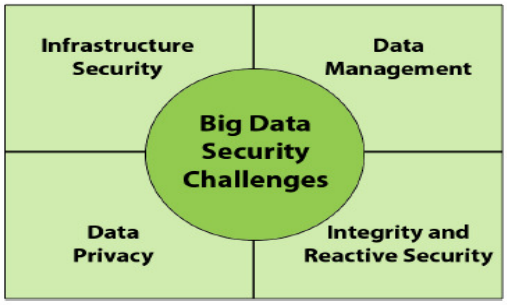

Fig 2: Security issues in Big data **[1]**

Because of the basic position of huge facts tools, the storage and processing of large volumes, these setup have now not yet reserved sufficient space to cover some applicable topics inclusive of security and privacy safety. Despite of this, the organization is too much on the benefits supplied from Big Data Analytics including Hadoop, and not so worried with privacy protection.

Rest of the paper is organized as follows, Section II contain the literature revive of security issues in hadoop Section III contain the Comparison of security approaches for Big-data Hadoop security with strength and limitations Section IV conclusion.

II. LITERATURE REVIEWS

Julio Moreno et. al.,[1] According to the working group of Big Data at the Cloud Security Alliance company there are, mainly, four distinctive elements of Big Data security: infrastructure safety, statistics privacy, statistics manipulate, and integrity and reactive protection. The fulfillment of the use of Big Data technology can be defined by way of using the discharge of a form of software program: Apache Hadoop. Hadoop is a structure developed through Apache that allows the designated preparing of huge datasets amid bunches of PCs utilizing programming styles. It is intended to be versatile from an unmarried server to piles of them, everything about offers calculation and near to carport. The frequently contracted arrangement as respects anchoring data privacy in a Big Data contraption is cryptography. Cryptography has been utilized to ensure data for a lot of time. Figuring on covered realities (CMD), which enhances records classification and honesty by utilizing allowing direct calculations to be made on veiled certainties, or new plans, comprising of Trusted Scheme for Hadoop Cluster (TSHC) which makes a fresh out of the box new design system for Hadoop a decent way to upgrade the privacy and wellbeing of the actualities. One of the greatest stretched out procedures wherein to secure the protection of data is by means of anonymize it. This comprises of making utilization of some state of approach or system to the insights which will put off the touchy certainties from it or to cover it. Huge Data normally infers a colossal measure of realities, and this inconvenience, accordingly, will increment in Big Data conditions.

Saraladevi et al.,[2] Kerberos is the community authentication protocol which allows the node to transfer any document over non secure channel by means of way of a device called rate price tag to expose their particular identification between them. This Kerberos mechanism is used to beautify the safety in HDFS. To growth the security in Hadoop base layers, the Bull eye Approach drastically utilized in HDFS to offer safety in 360 degree from node to node. This technique is brought on Hadoop module to view all sensitive facts in 360 degree to locate whether or not all the secured statistics are saved with none danger, and lets in the legal man or woman to keep the private data in a right manner. Nowadays this technique is using in groups like Data guise & DG comfortable and Amazon Elastic Map Reduce. In HDFS if there's any hassle in Name node event and will become unavailable, it makes the agency of device provider and statistics saved in the HDFS make unavailable so it isn't smooth to get proper of access to the facts in cozy way from this vital scenario. In order to boom the safety in information availability, it's far finished by means of the use of Name node.

Youssef Gahi et al.,[3] Presented the time period Big data and Big-data analytics, Highlighted the protection and privacy demanding situations taken into consideration by the usage of the Big-data tool and additionally stated about the possible solutions and strategies to relaxed the dispensed. He has discussed the subsequent possible answers to defend privacy in massive information: Rules and Legality, Encryption, Authentication, Metadata and Tagged Data, Unstructured distribution, Anonymization, Tracing interest.

R. R. Parmar et al.,[4] Proposed Data which is in rest alludes to measurements spared in relentless capacity. Of course Hadoop does not scramble data this is saved money on plate and that may reveal secure records to wellbeing assaults. This is particularly a vast problem because of the person of Hadoop structure, which spreads certainties all through a major scope of hubs; uncovering the data obstructs in any regard the ones unsecured access factors. Today, Hadoop isn't any additional a gadget for experimentation. Or maybe, it is getting developing notoriety at association level. It essentially offers those solid and cost compelling gigantic data stockpiling and handling stage and an obvious forceful advantage. Be that as it may, alongside that there are a few dangers related with it. For example, there can be a danger of information spillage in travel while it is getting exchanged over system from Hadoop shopper to Data Node. Proposed a brand new idea to eliminate the vulnerabilities within the Hadoop frame work known as 3D Security in Hadoop environ. This technique uses the Salsa20 and its variations like ChaCha20. In this techniques input record is split in to small chunks and people chunks are encrypted the usage of

Salsa20 and its variants like ChaCha20 earlier than send it to the Name Node in HDFS.

YANG Mengke et al., [5] Discussed New kinds and features of facts protection problems associated with big records and proposed feasible answer for the massive statistics security problems. Author discussed approximately the subsequent safety issues in the age of big information 1. Big information will growth the threat of statistics leakage: The danger on information leakage will result in the risk for non-public privacy.2.Big records grow to be apparent aim of cyber-assault: In cyber space, records base protected by means of manner of massive data is much less complicated to become the goal of hackers.3.Big information disturbing situations present keep and safety features: Big facts brings new protection annoying conditions for keep measures because of its variety. Big statistics used as an attacking degree: Hackers use big records generation to launch an attack to corporations whilst these companies use huge statistics to gain information for commercial price. The creator proposed the subsequent scheme for securing the large facts. 1. Establish big statistics safety manage platform: Data garage layer: at ease garage of the huge information can be accomplished thru encryption, 2nd is to divide secrete key and encrypted facts, 1/3 is to use filters, the transmission of the facts may be stopped as soon as statistics leaves the customers community, fourth is statistics again-up: using catastrophe tolerant machine centralized control of facts and statistics can be managed. Data processing layer: stream processing and batch processing. Interface: IAM (Identity and access control) is ready of commercial enterprise system and control measures for supplying inexperienced and cozy IT beneficial aid get admission to. Security of interface needs the aid of VPN encryption era and heritage linkage of managers together with some distance off lock and information wiping and automatic alarm.

Pradeep Adluru et al.,[6] Come up with a method to set up the accept as true with between the NameNode and the purchaser. User authentication to the NameNode is finished by means of Hashing technique (SHA-256). User authenticate by way of manner of sending the hash feature to NameNode the NameNode then generate the hash function observe with the hash characteristic sent by means of the person. Using RSA, Rijndael, AES and RC6, Randomly encrypt the statistics in the high-quality nodes of the cluster. In hadoop frame paintings the whole information is divided into chunks of 64MB, encryption and decryption of the records is dealt with via the MapReduce function of the Hadoop device.

Xianqing Yu, et al.,[7] Introduced a protection increased for an open cloud based Hadoop, named SEHadoop, to embellish the trade off strength through enhancing confinement among Hadoop parts and executing slightest access benefit for Hadoop procedures. He composed and performed SEHadoop form that comprises of SEHadoop runtime demonstrate, SEHadoop Block Token and SEHadoop Delegation Token to improve trade off strength of Hadoop in an open cloud.

Masoumeh RezaeiJam, et al.,[8] Discussed about the present day hadoop protection degree and also stated a few strategies like Apache Sentry for authorization, Fully Homomorphic Encryption, Authentication Using One Time PAD, Accessing HDFS Based On Attribute-Group, Triple Encryption Scheme For Hadoop-Based Data, Security Framework In G-Hadoop for robust authentication.

Yannan Maa et al.,[9] Proposed a Novel Approach for Improving Security and Storage Efficiency on HDFS. In order to enhance the storage efficiency, HDFS specifically advanced Hadoop Archive (HAR) to relieve this problem. Although, HAR occupies more memory usage and cannot replace statistics, at the equal time; the motive of limiting it's similarly development is that it does not drastically enhance the information transmission overall performance. Sequence record is confronted severe challenges in the same way as HAR. There are also strategies to enhance performance from device stage. The fundamental concept for WebGIS is to merge big files right into a big one and to construct index for every small record wherein hash index is used. Establishing secondary index listing or prefetching mechanism is likewise any other manner to improve analyzing performance. But each strategy above need high requirement for the integrity of the document stored. Network coding can be visible as a form of statistics encryption. The device cannot reason facts leakage or everlasting harm and exceptional unfavorable results in spite of the fact that parts or all of records inside the information lose or are stolen within the approach of transmission via the facts channel. He code the coding matrix twice all over again and store a part of the coding matrix storing the second one-encoding matrix and 0.33-encoding matrix within the datanode to enhance reading efficiency. It is crucial to be conscious that as a way to assure the safety of document garage, we can keep part of facts ultimately of every encoded encoding element in distinct backups, and the particular mapping for elements to factors may be saved in Namenode. The gain is that we make certain that each backup does now not incorporate an entire particular fact, even though the complete backup without a doubt lose, unlawful winner cannot restore any records correctly, due to the reality all encoding issue will not be complete. Unless the unlawful winner obtains the overpowering majority of a backup and extra wins the mapping for every element to component from Namenode at the same time, in any other case, it is absolute that there does not exist the opportunity of information being stolen.

Zichan Ruan, et al.,[10] Proposed the novel sampling approach, proof of the visualization set of regulations inside the KDD99 information set. The KDD99 information set as a robust benchmark since 1999. The benefit of proposed sampling technique is the manage of the inherent flaws that

massive records units include in phrases of volume, range and pace: Author implemented hash set of rules, a weight table, and sampling approach to deal with the inherent issues caused by analyzing large facts; extent, range, and pace. By make use of visualization algorithm, able to benefit insights into the KDD99 records set with a smooth identity of & everyday clusters and described tremendous clusters of powerful assaults.

Chao YANG, et al.,[11] With more and more cloud packages being to be had, facts safety becomes an vital issue in cloud computing. Data protection is a vital trouble in cloud computing environments. In HDFS all documents are saved in easy text and managed via a significant server. Thus, HDFS is not at ease toward storage servers that could peep at statistics content material fabric. Additionally, Hadoop and HDFS have a vulnerable security version; specifically the conversation amongst Data nodes and between customers and data nodes isn't always encrypted. So as to ensure measurements security in Hadoop-based cloud actualities carport, a solitary triple encryption conspire is proposed and actualized, which blends HDFS reports encryption utilizing DEA (Data Encryption Algorithm) and the insights scratch encryption with RSA, after which scrambles the individual's RSA private key utilizing IDEA (International Data Encryption Algorithm).

MATTURDI Bardi, et al,[12] Given definition of Big-facts: A broadly diagnosed definition belongs to IDC: information technology describe a modern day generation of generation and architectures, designed to economically extract fee from very big volumes of a giant fashion of facts, via permitting the excessive-speed capture, discovery, and/ or evolution. Author gives the insight approximately the present day safety and privacy issues of large statistics and additionally stated about a few exceptional answer available for protection and privacy troubles of large statistics: The maximum typically used choice to make sure safety and privacy can be oral and written legal tips, however the laws can't maintain tempo with the development of Technology and it is precise among nations. A novel technological named the included Rule-Oriented Data (iRODS) is proposed to be the answer to make certain safety and privacy in big information.

Madhvaraj M Shetty et al.,[13] With the amount of cloud packages is developing; the information safety becomes a chief problem in the cloud. In cloud computing environments, records protection is an essential trouble particularly even as the companies that manage touchy facts along with healthcare and monetary facts. The Hadoop is the one of the current fashion in generation that is used as a framework for cloud garage. But its miles designed without considering protection of facts saved, because it became a sincere device to keep and run jobs on large quantities of records. If there may be no proper facts safety, facts leakage is possibly to occur for the reason that intruders can gain statistics as plaintext without delay from the DataNode. At the bottom stage disk encryption can be provided via the going for walks device, this protects toward hard pressure theft. To defend statistics from unauthorized customers, Transparent Data Encryption (TDE) may be provided by HDFS. In TDE all equipment which hold their facts internal HDFS are able to take gain of it because it's miles carried out as an cease to give up answer because of this the statistics is encrypted and decrypted on the consumer side..

Thu Yein et al.,[14] Proposed a singular massive records primarily based completely safety analytics approach to detecting advanced assaults in virtualized infrastructures. Network logs as well as person software logs gathered periodically from the vacationer virtual machines (VMs) are saved inside the Hadoop Distributed File System (HDFS). Then, extraction of attacks abilities is achieved via graph-based occasion correlation and MapReduce parser primarily based identity of potential attack paths. Next, dedication of attack presence is carried out through-step device analyzing, particularly logistic regression is carried out to calculate Attacks conditional possibilities with respect to the attributes, and belief propagation is carried out to calculate the notion in life of an assault based totally on them. Experiments are done to assess the proposed method the usage of well-known malware in addition to in evaluation with present safety techniques for virtualized infrastructure.

Khairulliza Ahmad Salleha, et al.,[15] It's miles apparent that protection and privacy issues of massive records are not restrained to technological incapacity, in fact, the issues and demanding situations may additionally stand up from organizational tradition in addition to environmental factors. While the ones findings suggests the relevancy of searching at protection and privacy troubles of large statistics from exceptional perspectives, and specially how those troubles may play a function in encouraging/discouraging business huge facts adoption, it is but to be addressed empirically in IS publications. This truth opens up future research possibilities.

A RENCI/ National consortium for information technological know-how white paper., [16] Mentioned the annoying situations of statistics leakage and describes an revolutionary technological solution advanced thru RENCI team of workers; specifically the Secure Medical Workspace (SMW). Data leakage is a excellent challenge in virtual security and privacy due to the supply of huge quantities of facts, the functionality to link disparate records assets, the boom in facts sharing, and the absence of recommendations and tactics that mirror evolving technological competencies. The SMW addresses the challenge of facts leakage. It represents a primary reinforce over modern technological answers to the records leakage undertaking, which includes the incorporation of two-element authentication, DLP generation, virtualization generation, and organization-unique safety and privacy regulations that may be tailored or updated

as desired. The SMW is bendy and scalable and calls for little or no ongoing IT support.

C.L.Philip Chen et al., [17] Discuss a demanding situations of leakage of Big Data, which include Big Data packages, Big Data possibilities and stressful situations, as well as the extremely-current techniques and technology we presently undertake to cope with the Big Data issues, moreover communicate numerous underlying methodologies to address the information deluge, for example, granular computing, cloud computing, bio-stimulated computing, and quantum computing.

Tian, Y., [18] Big Data will assist to create new increase opportunities and actually new classes of organizations, which include those that combination and analyze company records. Meanwhile, with the momentum at the back of big records growing, a complete safety mechanism is wanted to mitigate threat of breach and assure the notable utilization of big facts technology, an outline of big information and its related security troubles are referred to. Also advocate a smart protection model for reaching large information safety.

Priya P. Sharma et al.,[19] Originally Hadoop modified into advanced without safety in thoughts, no safety model, no authentication of customers and offerings and no statistics privacy, so all people have to put up arbitrary code to be executed. Although auditing and authorization controls (HDFS file permissions and ACLs) were utilized in in advance distributions, such get admission to manipulate turned into without troubles avoided because any client may want to impersonate some other customer. Because impersonation become common and finished thru maximum users, the protection controls measures that did subsist had been no longer very effective. Later authorization and authentication grow to be brought, in any case, that to have some shortcoming in it. Since there were just a couple of security control measures inside Hadoop climate, numerous fortuity and assurance episodes occurred in such situations. All around expected clients should commit errors (e.g. Erasing enormous measures of records inside seconds with an apportioned erase). All customers and developers had the equivalent level of get right of section to benefits to every one of the measurements inside the bunch, any mission need to get admission to any of the data inside the group, and any client need to peruse any realities set. Since MapReduce had no understanding of verification or approval, a naughty client may need to bring down the needs of various Hadoop occupations with the aim to make his diversion whole faster or to be accomplished first-or more regrettable, he need to execute alternate employments.

Alfredo Cuzzocrea., [20] offer a top level view of state-of-the-art studies troubles and achievements in the subject of privacy and security of huge data, by means of highlighting open problems and actual research traits, and drawing novel studies directions in this field.

***Poonam R. Wagh et al., [21*]**, proposed system combination of security systems used to ensure security of the network to additional extent. A novel model of cloud secure storage is proposed, which combines the Hadoop distributed file system (HDFS) security protocol and cryptography for data stored on HDFS. The model uses the HDFS as the storage platform. Hadoop uses security protocols like Kerberos, LDAP and key tabs for user authentication. The fast encryption of cryptography algorithms and identity authentication like RSA, blowfish, AES can be used for overtime checking and the performance of Hadoop. Thus the security system with combination at different levels can supply secured, effective, stable effect on large data through Hadoop and cloud computing.

Anitya Kumar Gupta et al.,[22] Examine security issues for cloud computing, Big Data, Map Reduce and Hadoop environment. The fundamental focus is on security issues in distributed computing that are connected with huge information. Big data information applications are an awesome advantage to associations, business, organizations and numerous substantial scale and little scale ventures. Author additionally talks about different conceivable answers for the issues in distributed computing security and Hadoop. Distributed computing security is creating at a quick pace which incorporates PC security, system security, data security, and information protection. Distributed computing assumes an extremely indispensable part in securing information, applications and the related foundation with the assistance of arrangements, innovations, controls, and enormous information instruments. In addition, distributed computing, enormous information and its applications, focal points are prone to speak to the most promising new frontiers in science.

Table 1: Comparison of security approaches with strength and limitations.

Sl no	Year	Author	Working Characteristics	Strength	Limitations
[1]	2017	R. R. Parmar et al.,	Three-Dimensional Security in Hadoop Environment.	Map-Reduce tasks can directly be executed on data stored in HDFS without requiring the complete file to be decrypted before it can be processed and hence it completely eliminates the use of HDFS encryption zone.	Individual data blocks stored in Map-Reduce are needed to be decrypted at the time of Mapping.
[2]	2015	Pradeep Adluru et al.,	SHA-256,RSA, Rijndael,	Even if the hacker manages to break into a single cluster or node on the cloud and manages to	One of the primary issues of Big Data is velocity and so time is

			AES,RC6	break the cipher text, He will not gain access to the entire data. Without access to the entire data, the hacker will not be able to process the data for any information.	the biggest concern as it might lead to loss of information if the system is occupied in encrypting the information instead of collecting the information.
[3]	2014	Xianqing Yu, et al.,	SEHadoop runtime model SEHadoop Block Token and SEHadoop Delegation Token	By implementing SEHadoop, compromising small parts of Hadoop only has limited impact on the rest of components inHadoop, therefore SEHadoop improves compromise resilience of Hadoop in a public cloud.	If attackers can launch a large scale or system wide attacks to Hadoop in a public cloud, e.g. compromising cloud management system, SEHadoop only has limited benefits.
[4]	2015	Yannan Maa et al,	Network Coding and Multi-node reading	The ability of security and storage efficiency of the system through special decoding mode and multi-nodes reading greatly improved. Meanwhile, we can store part of the metadata in Datanode to reduce the memory usage of the system, so that the storage efficiency of the system may improve obviously through the decrease of Namenode's workload because work of data encoding is done on the Namenode.	In this approach, file storage time was extended by about 25%, because files needed to be encoded before storing.
[5]	2017	Zichan Ruan, et al,	Hash algorithm, a weight table, and sampling method	Proposed sampling method is the management of the inherent flaws that big data sets comprise in terms of volume, variety and velocity. By exploiting the hash algorithm, weight table, and weight equation, addressed the bias and redundancy problems caused by volume and variety of the KDD99 data set.	The clusters of several attack types remain overlapped. Further study focusing on inter-class distinctness among those attack types could provide information and underlying relationship of those classes.
[6]	2013	Chao YANG, et al,	DEA,RSA,IDEA	HDFS files are encrypted by using the hybrid encryption based on DES and RSA, and the user's RSA private key is encrypted using IDEA. The triple encryption scheme is implemented and integrated in Hadoop-based cloud data storage.	The parallel processing of the encryption and decryption using Map-Reduce, in order to improve the performance of data encryption and decryption can be adopted
[7]	2018	Thu Yein et al,	Two-step machine learning, namely logistic regression and belief propagation	Detecting both botnets and in-VM malware. Real-time automated retraining of classifiers.	Occasional latency increase due to SSH server reset by guest VMs
[8]	2013	A RENCI/ National consortium for data science white paper	Secure Medical Workspace (SMW).	The SMW addresses the challenge of data leakage. It represents a major advance over current technological solutions to the data leakage challenge, including the incorporation of two factor authentication, DLP technology, virtualization technology, and institution-specific Security and privacy policies that can be tailored or updated as needed.	Enhanced, open source version of the SMW architecture that incorporates Hypervisor-based, distributed DLP technology.
[9]	2017	Tian, Y	Intelligence driven security model for big data.	Review the current data security in big data and analysis its feasibilities and obstacles. Introduced intelligent analytics to enhance security with the proposed security intelligence model.	

III. COMPARISION TABLE

The Table I. shows the comparison among various Techniques. With the help of comparison table noticed that Encryption is one of the high priority techniques for Big-data Hadoop security. However, Encryption causes execution corruption and this must be deliberately assessed in future, by scrambling imperative delicate information to be anchored rather encoding every one of the information. In Future, there is need of creating form of Hadoop with an extensive variety of security methods for anchoring data and moreover secure execution of occupation inside the system.

IV. CONCLUSION

The clear trend in Big Data towards high rate of collection, storage, and linkage of personal data sets raises security concerns relating to theft, institutional misuse, and re-identification, as well as unauthorized access, redistribution, and repurposing of personal information. In order to save you these concerns from turning into real harms, effective coverage and technological measures are required on the a part of corporations that use "Big Data", as well as for individuals to whom the data relates. We have seen that the pleasant answer of enforcing Big Data protection and

privacy is the law rather than the safety technology however the laws can't maintain tempo with the improvement of technology and is different between countries. Thus, protection technology and other techniques are constantly necessary. In Big Data Era, where statistics is gathered from various sources, protection is a main challenge (essential requirement) as there may be no fixed supply of information. With the Hadoop gaining large recognition in the industry, a natural difficulty over the security has spread. A growing want to accept and assimilate these security answer and industrial security features has surfaced. In this paper, we've highlighted a fixed of safety and privacy demanding situations that have to be taken into consideration by means of huge records equipment. Furthermore, we supplied some feasible answers and techniques that would assist securing this allotted environment.

REFERENCES

[1] Julio Moreno, Manuel A. Serrano and Eduardo Fernandez-Medina "Article Main Issues in Big Data Security" Future Internet 2016, 8, 44; doi: 10.3390/fi8030044.

[2] B. Saraladevi, N. Pazhaniraja, P. Victer Paul, M.S. Saleem Basha, P. Dhavachelvan "Big Data and Hadoop-A Study in Security Perspective" 2nd International Symposium on Big Data and Cloud Computing (ISBCC'15) doi: 10.1016/j.procs.2015.04.091

[3] Youssef Gahi, Mouhcine Guennoun, and Hussein T. Mouftah "Big Data Analytics: Security and Privacy Challenges" IEEE Symposium on Computers and Communication (ISCC), 2016

[4] Raj R. Parmar, Sudipta Roy, Debnath Bhattacharyya,Samir Kumar Bandyopadhyay, (Senior Member, IEEE), And Tai-Hoon Kim "Large-Scale Encryption in the Hadoop Environment: Challenges and Solutions" IEEE Access, 2017. Digital Object Identifier 10.1109/ACCESS.2017.2700228

[5] YANG Mengke, ZHOU Xiaoguang, ZENG Jianqiu, XU Jianjian, "Challenges and solutions of information Security issues in the Age of Big Data" China communications, 2016

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

[7] Xianqing Yu, Peng Ning, Mladen A. Vouk, "Enhancing Security of Hadoop in a Public Cloud" 6th International Conference on Information and Communication Systems (ICICS), IEEE 2015

[8] Masoumeh RezaeiJam, Leili Mohammad Khanli, Mohammad Kazem Akbari, "A Survey on Security of Hadoop" IEEE 4th international conference on computer and knowledge engineering, 2014

[9] Yannan Maa, Yu Zhoua, Yao Yua, Chenglei Penga, Ziqiang Wanga, Sidan Dua "A Novel Approach for Improving Security and Storage Efficiency on HDFS" The 6th International Conference on Ambient Systems, Networks and Technologies (ANT 2015) doi: 10.1016/j.procs.2015.05.062

[10] Zichan Ruan , Yuantian Miao, Lei Pan, Nicholas Patterson, Jun Zhang "Visualization of big data security: A Case study on the KDD99 cup data set" Digital Communications and Networks 3 (2017) 250–259 doi.org/10.1016/j.dcan.2017.07.004

[11] Chao YANG, Weiwei LIN, Mingqi LIU "A Novel Triple Encryption Scheme for Hadoop-based Cloud Data Security" Fourth International Conference on Emerging Intelligent Data and Web Technologies, IEEE 2013 DOI 10.1109/EIDWT.2013.80

[12] MATTURDI Bardi, ZHOU Xianwei, LI Shuai, LIN Fuhong "Big Data security and privacy: A review",China Communications Supplement No.2, 2014

[13] Madhvaraj M Shetty, Manjaiah D.H "Data Security in Hadoop Distributed File System" International Conference on Emerging Technological Trends [ICETT], IEEE 2016

[14] Thu Yein Win,Member IEEE, Huaglory Tianfield, and Quentin Mair, Member IEEE "Big Data Based Security Analytics for Protecting Virtualized Infrastructures in Cloud Computing" IEEE TRANSACTIONS ON BIG DATA, VOL. 4, NO. 1, JANUARY-MARCH 2018

[15] Khairulliza Ahmad Salleha,Lech Janczewski "Technological, organizational and environmental security and privacy issues of big data: A literature review" Procedia Computer Science 100 (2016) 19 – 28, doi: 10.1016/j.procs.2016.09.119

[16] A RENCI/ National consortium for data science white paper "Security and Privacy in the Era of Big Data" November 2013

[17] C.L.Philip Chen,Chun-Yang Zhang "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data" Information Sciences 275 (2014) 314–347, dx.doi.org/10.1016/j.ins.2014.01.015

[18] Tian, Y. (2017) "Towards the Development of Best Data Security for Big Data".Communications and Network, 9, 291-301.

[19] Priya P. Sharma, Chandrakant P. Navdeti "Securing Big Data Hadoop: A Review of Security Issues, Threats and Solution" International Journal of Computer Science and Information Technologies, Vol. 5 (2), 2014, 2126-2131

[20] Alfredo Cuzzocrea "Privacy and Security of Big Data: Current Challenges and Future Research Perspectives" Copyright © 2014 ACM 978-1-4503-1583- 8/14/11

[21] Poonam R. Wagh, Amol D. Potgantwar "Providing Security to Data Stored on HDFS Using Security Protocol" International Journal of Scientific Research in Network Security and Communication Volume-5, Issue-4, August 2017, ISSN: 2321-3256

[22] Anitya Kumar Gupta, Srishti Gupta "Security Issues in Big Data with Cloud Computing" International Journal of Scientific Research in Computer Sciences and Engineering, Vol.5, Issue.6, pp.27-32, December (2017) E-ISSN: 2320-7639

[23] Masoumeh RezaeiJam, Leili Mohammad Khanli, Mohammad Kazem Akbari "A Survey on Security of Hadoop",4th international conference on computer and knowledge engineering (ICCKE) IEEE 2014.

Table 2: Current security works in Big Data Hadoop

Reference	Context of research	Problem Discussed	Technique Used	Model/ Tool proposed
[1]	Big data security Issues	Main problems and challenges related to security in Big Data.	Theory	NO
[2]	Big Data and Hadoop-Security Perspective	Focused more on security issue arises in Hadoop Architecture base layer called Hadoop Distributed File System (HDFS).	Kerberos, Bull Eye Algorithm and Name node.	NO

[3]	Security and Privacy Challenges of Big Data Analytics.	Benefits of Big Data Analytics and review on challenges of security and privacy in big data environments.	Theory	NO
[4]	Security issues associated with Hadoop	Vulnerabilities in the Hadoop	Three-Dimensional Security in Hadoop Environment.	YES
[5]	Information security Issues in the age of Big data.	Main problems and challenges related to security in Big Data	Theory	NO
[6]	Hadoop Security	Hadoop and difficulty in maintaining the privacy and security of BigData.	SHA-256,RSA, Rijndael, AES,RC6	YES
[7]	Enhancing Hadoop Security in public cloud.	Potential internal cloud attacks to Hadoop in a public cloud.	SEHadoop runtime model SEHadoop Block Token and SEHadoop Delegation Token	YES
[8]	Hadoop Security	Current hadoop security level	Apache Sentry, Fully-Homomorphic Encryption, Authentication Using One Time PAD, Accessing HDFS Based On Attribute-Group, Triple Encryption Scheme For Hadoop-Based Data, Security Framework In G-Hadoop	NO
[9]	Security and Storage Efficiency on HDFS	Hadoop Archive (HAR) WebGIS	Network Coding and Multi-node reading	YES
[10]	Visualization of big data security, on the KDD99 cup data set	Intrusion detection systems (IDSs).	Hash algorithm, a weight table, and sampling method	YES
[11]	Triple Encryption Scheme for Hadoop-based Cloud Data	Data security protection in the Cloud. To ensure data security in cloud data storage.	DEA,RSA,IDEA	YES
[12]	Big Data security and privacy	The insight about the current security and privacy issues of big data and also discussed about some best solution available for security and privacy of big data.	Integrated Rule-Oriented Data (iRODS)	NO
[13]	Describe present security in Hadoop Distributed File System.	Ensure the security of important data at an HDFS storage level which has not been achieved by Kerberos.	Theory	NO
[14]	Detecting advanced attacks in virtualized infrastructures.	Big data based security analytics (BDSA) approach to detecting advanced attacks in virtualized infrastructures.	Two-step machine learning, namely logistic regression and belief propagation	YES
[15]	Technological, organizational and environmental security and privacy issues of big data	Security and privacy issues of big data	Theory	NO
[16]	Security and Privacy in the Era of Big Data	Challenges of data leakage	Secure Medical Workspace (SMW).	YES
[17]	Data intensive scientific discovery (DISD), also known as Big Data problems	Big Data applications, Big Data opportunities and challenges, state-of-the-art techniques and technologies we currently adopt to deal with the Big Data problems	Theory	NO
[18]	Development of Best Data Security for Big Data	Current data security in big data and analysis its feasibilities and obstacles	Intelligence driven security model for big data.	YES

[19]	Securing Big Data Hadoop	Security solution to secure the Hadoop ecosystem.	Theory	NO
[20]	Privacy and Security of Big Data	Overview of state-of-the-art research issues and Achievements in the field of privacy and security of big data, by Highlighting open problems and actual research trends, and drawing novel research directions in this field.	Theory	NO

Authors Profile

Dr K C Ravishankar Professor in Department of Computer Science and Engineering, Government Engineering College, B M Road, Dairy Circle, HASSAN. He obtained his B.E. from Mysuru University Mysuru, M.Tech from Indian Institute of Technology Delhi, and Ph.D from Vishweswaraya Technological University Belgaum, Karnataka. His research areas include the Information and Network Security, Big data, Cloud computing and image processing.

Mr. Sathisha M S, Asst. Professor, Faculty in Department of Computer Science and engineering, Canara Engineering College, Mangalurure, He obtained his B.E and M.Tech from Vishweswaraya Technological University Belgaum, Karnataka in 2005 and 2011 respectively, currently pursuing Ph.D in Vishweswarya Technological University Belgaum, Karnataka. His research areas include the Information and Network Security, Big data and Cloud computing.

