iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** IV **Month of publication:** April 2022

DOI: https://doi.org/10.22214/ijraset.2022.41155

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue IV Apr 2022- Available at www.ijraset.com

A look at the Security and Privacy of Big Data

Harshit Poddar
School of Electronics Engineering (SENSE), Vellore Institute of Technology, Vellore, India

Abstract: ***In recent years a lot of data is being generated every second. So much data is produced that our conventional data management tool is not enough for this. The data is growing day by day because of more uses of the internet as well as more users of the smartphone, social media networks, Health care, etc. So, because of this, we made a term known as Big Data. So, from above we can understand that big data is a collection of a very large amount of data as well as complex data. Data is generally stored in three forms are structured, unstructured, semi-structured data. To answer these questions, we will perform a small rule which is called a "systematic Literature review (SLR)" [1]. According to this rule, we will collect, categorize, analyze and give the conclusion from the top research papers, review papers, books, conferences, etc. The main contribution of this work lies in the comparative study of big data security and privacy. We will try to find the issues of privacy and security in big data but still, we require some more research in this field.***

Keywords: Big data, privacy, security, Hadoop framework, Volume, Velocity, Variety, Veracity, Vocabulary, Vagueness, Viability, Value, Database.

I. INTRODUCTION

The term big data is now generally used everywhere in our daily life. Big data is a collection of huge data in volume as well as more complex data which is growing exponentially with time. The data is so big that our conventional data management tools are not sufficient for this humongous data. Like in social media 500+terabytes of new data are given to the database of social media every day. This is just an example of social media. If we took all the data of health care, insurance, office data, etc then you can't imagine how big is big data. The data is generally stored in three forms: a) Structured: Structured data is any type of data that can be stored, retrieved, processed in a fixed form. b) Unstructured: Unstructured data is information that is not organized in a pre-defined manner or does not have a pre-defined data model. c) Semi-structured: Semi-structured data is information that doesn't consist of Structured data (relational database) but still has some structure in it.

Figure 1: Structured, Semi-structured, Unstructured

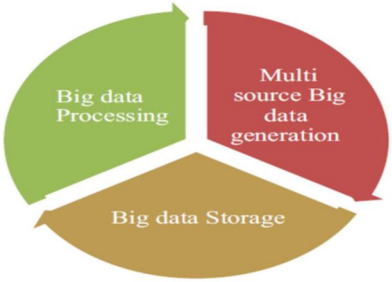


Figure 2: "Big data life cycle stages of the big data life cycle, i.e., data generation, storage, and processing is shown" [4].

This paper will mainly focus on big data security and privacy and we will try to find some issues and try to work on the solution for these issues.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue IV Apr 2022- Available at www.ijraset.com

II. BIG DATA CHARACTERISTICS: THE 8V'S

So above we discussed big data and some different types of data in which they are stored. Now here we will discuss the characteristics of big data. As characteristics of big data are not is not completely defined as some are saying there are 5V's, some 10V's, and many more. But here we will take in general therefore 8V's. So, the 8V's are as follows: -

1) *Volume:* From the word itself we can understand that volume is the amount of data needed to be processed at a given time. This is the basic definition of volume but according to big data volume will have some size limit therefore if the volume is extremely big like petabytes, terabytes, exabytes, etc then we consider it as big data.

Table 1: Estimated data generated

Company Name	Estimated Data generated
UPS	16 petabytes (per Day)
Walmart	2.5 petabytes (per hour)
Facebook	100 terabytes (daily)
Shell	10 Exabyte (annually)

2) *Velocity:* This is related to the speed of the data going in and the speed of the converted data exiting the compute, just like Volume. Telemetry that must be evaluated in real-stime for a self-driving automobile is an example of a high-velocity need.
3) *Variety:* As we know that big data deals with any data format – structured, semi-structured, unstructured, or even more complex structured data. So, storing and processing unformatted data through Relational Database Management System (RDBMS) is not so easy. Unstructured data, on the other hand, gives more important insights into the information that structured data seldom delivers. Furthermore, a variety of data implies a range of data sources. As a result, this big data feature also gives information about the data sources.
4) *Veracity:* As we know that not all the data that come for the processing is valuable. So, unless we do not clean the data correctly, it is not advisable to store or process complete data. As the data volume is humongous, so from here another "V" comes which is veracity. This characteristic also helps to determine whether the data is coming from a legit source or not and whether it is a right fit for the analytic model as well as this will save our precious time for not processing invaluable data.
5) *Value:* This is the most important "V" from all of them. Big data is most likely being pursued its business worth. This is maybe the most important aspect of big data. Because other big data features have no relevance until you gain business insights from them.
6) *Visualization:* There is no use in evaluating anything unless it is expressed or shown in a meaningful way. As a result, large data must be displayed using a proper tool that serves different parameters to assist data scientists or analysts in better understanding it. Plotting billions, millions of data points is a difficult operation. Therefore, it associates with other methodologies such as employing tree maps, network diagrams, cone trees, and so on.
7) *Vulnerability:* This is also an important "V" on which we should think about it. As most of the big data resources are open source, there is a possibility that hackers can attack them. h) Variability: In the context of big data, variability may relate to a number of distinct things. The number of discrepancies in the data is one. In order for any useful analytics to take place, they must be discovered using anomaly and outlier detection tools. Big data is also changeable due to the plethora of data dimensions arising from a variety of distinct data kinds and sources. Variability may also relate to the inconsistency with which large amounts of data are fed into your database.

III. DIFFERENCE BETWEEN DATA PRIVACY AND DATA SECURITY IN BIG DATA.

A. Data Privacy

privacy refers to an individual's right to be free from interference and prying eyes, or the right to be left alone. In many industrialized nations, it is protected by the constitution, making it a fundamental human right and one of the key concepts of human dignity, an idea on which most people can agree.

B. Data Security

Data security is totally focused on protecting personal data from any unauthorized third-party access or some unwanted attacks and exploitation of data. Data security methods practices are as follows

1) Activity monitoring
2) Breach response
3) Encryption
4) Network security

Above are some practices are taken to secure our data.

S.No	Privacy	Security
1	Privacy is the appropriate use of user's information	Security is the "confidentiality, integrity and availability" of data
2	Privacy is the ability to decide what information of an individual goes where	Security offers the ability to be confident that decisions are respected
3	The issue of privacy is one that often applies to a consumer's right to safeguard their information from any other parties	Security may provide for confidentiality. The overall goal of most security system is to protect an enterprise or agency [72]
4	It is possible to have poor privacy and good security practices	However, it is difficult to have good privacy practices without a good data security program
5	For example, if user make a purchase from XYZ Company and provide them payment [13] and address information in order for them to ship the product, they cannot then sell user's information to a third party without prior consent to user	The company XYZ uses various techniques (Encryption, Firewall) in order to prevent data compromise from technology or vulnerabilities in the network

Figure 3: Difference between Privacy and Security [4].

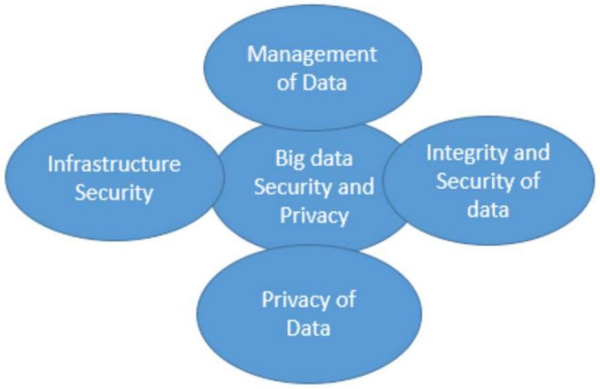


Figure 4: "Big data security and privacy" [5].

IV. ISSUES AND CHALLENGES IN BIG DATA SECURITY AND PRIVACY

We know that big data is an emerging field but we still require a lot of research and development as well as advancement in this field. As data is increasing day by day, we also need to be concern more about the privacy and security of our data. So here we have read, reviewed, researched some issues from different sources, and note down some important issues which we will be explaining below.

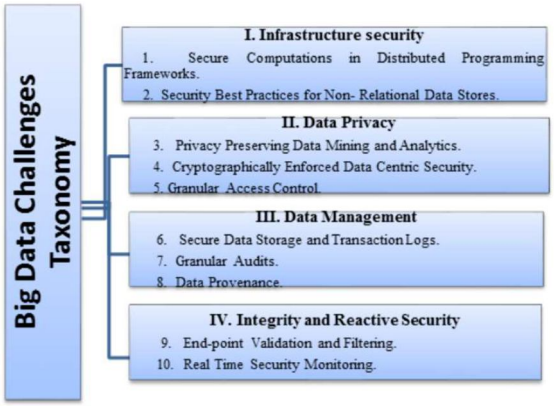


Figure 5: "Taxonomy of Top Big Data Challenges" [3].

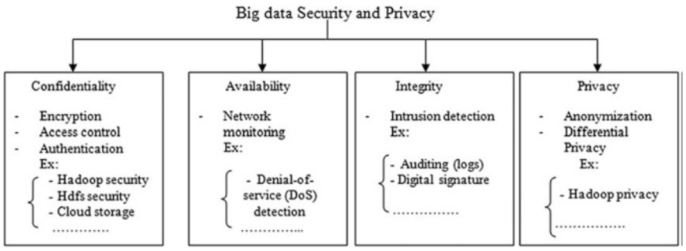


Figure 6: "Big data for security and privacy approaches" [6].

1) Nowadays, real-time security monitoring is on boom. But here is a big challenge or issue in this. When the real-time security monitoring devices make an alert on the basis of some situation then that alert is transferred to the security devices. So, in these alerts some alerts can be not useful and may lead to "false positives and due to human incapability to successfully deal with such a huge amount of them at such a speed, results in them being clicked away or ignored" [3].

2) In today's world internet banking, cryptocurrency and other online modes of transaction happen in a huge amount every second. So, in this data and transaction logs are stored in "multi-tiered storage media manually moving the data between tiers gives the its manager direct control over exactly what data is moved and when. However, as the size of the data set has been and continues to be, growing exponentially, scalability and availability have necessitated auto-tiring for big data storage management. Auto-tiring solutions do not keep track of where the data is stored, which poses new challenges to secure data storage" [3].

3) "Granular auditing and access control which is provided by databases like NoSQL or Hadoop Distributed File System necessitate a very robust process of authentication and compulsory access control" [5].

4) There should be efficient handling of the big data stream. Some specific scenarios are "stock exchange would require analysis of data in the form of stream. Fast and optimized solutions should be developed to make inference on big data a stream." [7].

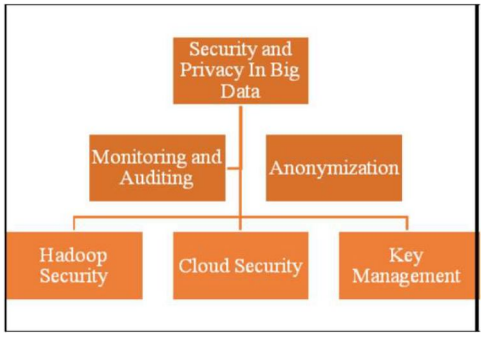


Figure 7: Categorization of security and privacy [5].

V. SECURITY AND PRIVACY IN BIG DATA: A SOLUTION.

The general solution or the procedure in security and privacy in big data is:

Figure 8: "Current system of Data Encryption" [2].

The above figure tells us about the general data encryption process but now a group of researchers has proposed a new system of data encryption that is more efficient as well as more secure than our conventional data encryption process.

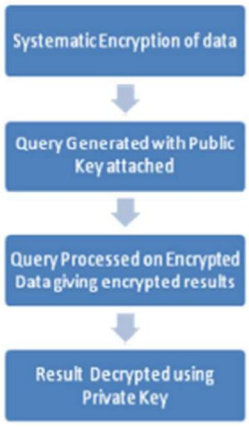


Figure 9: Proposed System of Data Encryption [2]

"Access control technology: Due to the huge no of users and complex authority in the big data environment, new technology must be adopted to realize the controlled sharing of data. The Role based access control (RBAC) is widely used access control mode" [3]. We should also try to enhance us encryption process as that is the weakest part and favourite part for hackers. Most of the hackers try to find some loophole in the encryption process and try to leak some data. So here we can use Homomorphic Encryption Schemes (HES). Usage of Hybrid clouds Hybrid cloud is a cloud computing environment that utilizes a blend of on-premises, private cloud, and third-party, public cloud services with the organization between the two platforms [4].

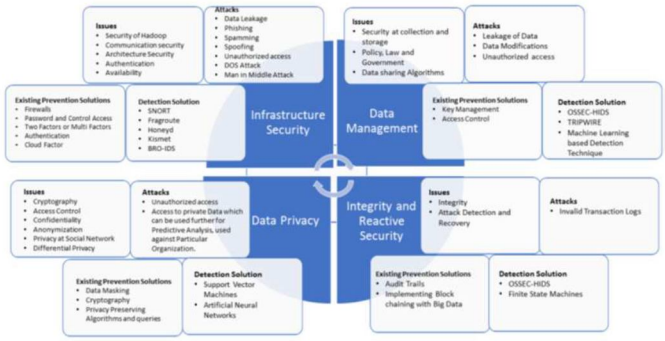


Figure 10: "Security and Privacy issues and attacks along with solutions in Big data" [5].

We can also use Data Anonymization / Deidentification technique. This technique is really good. In these techniques, the private and sensitive data is preserved which is used when big data is published to third parties. "Normally a record in a dataset consists of 3 types of attributes:

- Key attributes are those attribute that uniquely identifies each individual. e.g., ID, Name, address, phone number. They are always removed before release.
- Quasi-Identifiers (QI) are those sets of attributes that can be linked with other datasets which are publicly available to identify an individual's private data. It can be used for linking anonymized datasets with other datasets. e.g., Age, sex, zip code, city, etc.

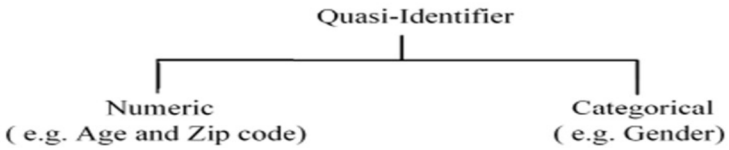


Figure 11: Types of QI attributes [11].

Sensitive attributes contain some sensitive information that an individual wants to hide from others. e.g., income, salary, disease, medical records, etc" [3].

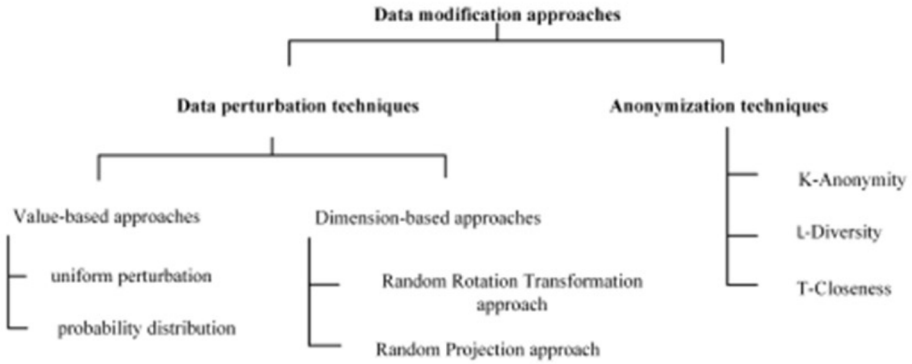


Figure 12: "Classification framework of the data modification approaches" [11].

There are three privacy-preserving methods of data anonymization that helps to prevent attacks on the privacy of the published data. The three methods are as follows: -

1) *K-anonymity:* When attributes are suppressed or generalized until each row is identical with at least k-1 other rows then that method is called a k-anonymity. It prevents definite database linkages and also guarantees that the data released is accurate. But it has some limitations:
   - It does not hide individual identity. Unable to protect against attacks based on background knowledge.
   - K-anonymity cannot be applied to high dimensional data" [3].

2) *L-diversity:* By decreasing the granularity of a data representation, l-diversity is a type of group-based anonymization that is used to maintain privacy in data sets.
3) *T-closeness:* This technique protects against homogeneity and background knowledge assaults while maintaining privacy. "The difference between the distribution of a sensitive property in the same class and the distribution of the attribute in the entire table is referred to as t-closeness when it is less than a certain threshold. If all equivalence classes in a table have t-closeness, the table is said to have t-closeness" [3].

Anonymization Approach	Advantage	Disadvantages
K-Anonymity Approach	Data remains truthful	Prone to background knowledge attack and Homogeneity attack
L-diversity Approach	Data remains truthful	Prone to similarity and skewness attack
T-Closeness Approach	Data remains truthful	Excessive information loss

Figure 13: A summary of data anonymization approaches [11].

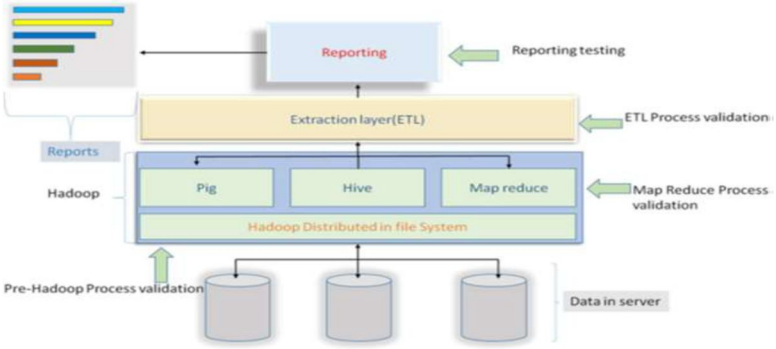


Figure 14: "Big data architecture and testing area new paradigms for privacy conformance testing to the four areas of the ETL (Extract, Transform, and Load) processes are shown here" [4].

COMPANY NAME	APPLICATION	KEY FEATURES
IBM	IBM QRadar Security Intelligence Platform [13]	• A comprehensive, integrated approach that combines real time correlation for continuous insight, custom analytics across massive structured and unstructured data and, forensic capabilities for deep visibility. The entire combination can help to address advanced persistent threats, fraud and insider threats. • High-speed querying of security intelligence data. • Graphical front-end tool for visualizing and exploring big data.
INFOSYS	IIP-The Infosys Information Platform [14]	• An open source data analytics platform. • Enables businesses to operationalize their data assets and uncover new opportunities for rapid innovation and growth. • Provides an end-to-end data platform that leverages open source innovations and internal enhancements to seamlessly integrate into enterprise landscapes in a way that it can operate as a standalone big data solution or as an add-on to existing proprietary tools.
MICROSOFT	Big Data and Business Intelligence Solutions [15]	• Provides a modern data management layer that supports all data types of data i. e. structured semi-structured and unstructured data at rest or in motion. MS makes it easier to integrate, manage and present real-time data streams, providing a more holistic view of business to drive rapid decisions. • An enrichment layer that enhances our data through discovery, combining with the world's data and by refining with advanced analytics. • An insights layer that provides insights to all users through familiar tools like Office's Excel & PowerPoint. • HDInsight, MS's new Hadoop based service that offers 100% compatibility with Apache Hadoop. It enables the customers to gain business insights from variety of data with any size and activate new types of data irrespective of its location.
HP	HPE Security- Hewlett Packard Enterprise [16]	• Security provides best-in-class data encryption and tokenization for structured and unstructured data. • Cost-effective PCI compliance, scope reduction, and secure analytics. • Used by leading companies worldwide, reducing risk and protecting brand.
ORACLE	DBSAT- The Oracle Database Security Assessment Tool [17]	• Quickly identify security configuration errors in the databases. • Promote security best practices. • Improve the security posture of Oracle Databases. • Reduce the attack surface and exposure to risk.
VORMETRIC	Vormetric Data Security Platform [18]	• Enable companies to maximize the benefits of big data analytics. It offers the granular controls, robust encryption, and comprehensive coverage that organizations need to secure sensitive data across their big data environments • Enables security teams to leverage centralized controls that optimize efficiency & compliance adherence. • It offers capabilities for big data encryption, key management and access control.

Figure 15: "Top companies Big Data Security Solution Applications and their Key Features" [3].

VI. CONCLUSION

Big data has proved himself that this can be the best way to analyze humongous and complex data. In big data, our topmost priority will be the privacy and security of our data as there will be many sensitive data which is really important to get encrypted. In this paper, we have discussed and analyzed various papers, books, and other important resources and we have found the important issues, challenges, and solutions in big data. This is really a good field but still, a lot of research is required for the advancement and increasing the efficiency in big data.

REFERENCES

[1] 2016 IEEE International Conference on Big Data (Big Data) 978-1-4673-9005-7/16/$31.00 ©2016 IEEE 3693 Security and Privacy for Big Data: A Systematic Literature Review Boel Nelson Department of Computer Science and Engineering Chalmers University of Technology Email: boeln@chalmers.se Tomas Olovsson Department of Computer Science and Engineering Chalmers University of Technology Email: tomasol@chalmers.se

[2] International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163 Issue 2, Volume 2 (February 2015) www.ijirae.com 2015, IJIRAE- All Rights Reserved Page – 15 Big Data Security Issues and Challenges Raghav Toshniwal* Kanishka Ghosh Dastidar Asoke Nath Department of Computer Science Department of Computer Science Department of Computer Science St. Xavier's College (Autonomous) St. Xavier's College (Autonomous) St.Xavier's College (Autonomous) Kolkata, India Kolkata, India Kolkata, India

[3] International Journal of Computer Applications (0975 – 8887) Volume 177 – No.4, November 2017 23 Big Data Security and Privacy: A Review on Issues, Challenges and Privacy Preserving Methods, Anupama Jha Assistant Professor, VIPS, GGSIPU New Delhi, India Meenu Dave, PhD Professor Jagannath University Jaipur, India Supriya Madan, PhD Professor VIPS, GGSIPU New Delhi, India

[4] Jain et al. J Big Data (2016) 3:25 -DOI 10.1186/s40537-016-0059-y big data privacy: a technological perspective and review Priyank Jain*, Manasi Gyanchandani and Nilay Khare

[5] 2021 3rd International Conference on Signal Processing and Communication (ICPSC) | 13 –14 May 2021 | Coimbatore A Review on Big Data: Privacy and Security Challenges, Parth Goel Department of Computer Science & Engineering Devang Patel Institute of Advance Technology and Research (DEPSTAR) Charotar University of Science and Technology (CHARUSAT), CHARUSAT campus, Changa 388421, India, Radhika Patel Department of Information Technology Devang Patel Institute of Advance Technology and Research (DEPSTAR) Charotar University of Science and Technology (CHARUSAT), CHARUSAT campus, Changa 388421, India, Dweepna Garg Department of Computer Engineering Devang Patel Institute of Advance Technology and Research (DEPSTAR), Charotar University of Science and Technology (CHARUSAT), HARUSAT campus, Changa 388421, India Amit Ganatra Department of Computer Engineering Devang Patel Institute of Advance Technology and Research (DEPSTAR) Charotar University of Science and Technology (CHARUSAT), CHARUSAT campus, Changa 388421, India, amitganatra.ce@charusat.ac.in

[6] A Comparative Study of Recent Advances in Big Data for Security and Privacy Ahlam Kourid and Salim Chikhi A. Kourid · S. Chikhi Computer Science Department, MISC Laboratory, College of NTIC, Constantine 2 University – A. Mehri, 25000 Constantine, Algeriae-mail:ahlam.kourid@univ-constantine2.dz S. Chikhi © Springer Nature Singapore Pte Ltd. 2018 G.M. Perez et al. (eds.), Networking Communication and Data Knowledge Engineering, Lecture Notes on Data Engineering and Communications Technologies 4, https://doi.org/10.1007/978-981-10-4600-1_23 BIG DATA, CLOUD

& MOBILE COMPUTING Big Data security and privacy: A review MATTURDI Bardi1, ZHOU Xianwei2, LI Shuai2, LIN Fuhong2* 1 School of mathematics and information, Hotan Teachers College, Hetian 848000, Xinjiang, P. R. China 2 School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, P. R. China

[7] Real-Time Processing of Big Data Streams: Lifecycle, Tools, Tasks, and Challenges Fatih Gürcan Department of Computer Engineering Karadeniz Technical University Trabzon, Turkey fgurcan@ktu.edu.tr Muhammet Berigel Department of Management Information Systems Karadeniz Technical University Trabzon, Turkey berigel@ktu.edu.tr

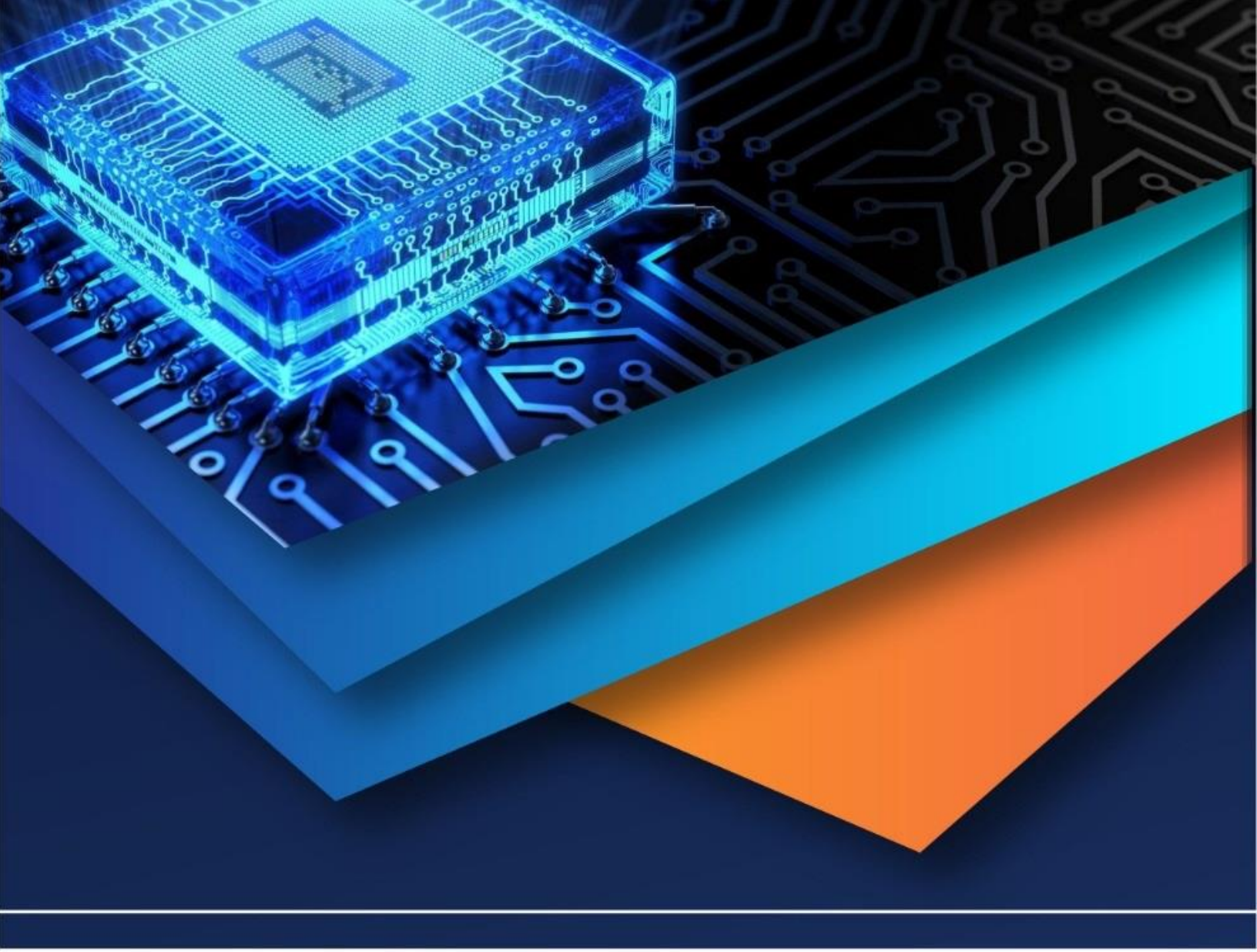
[8] Contents lists available at ScienceDirect Array Journal Big Data: Hadoop framework vulnerabilities, security issues and attacks Gurjit Singh Bhathal *, Amardeep Singh a department of Computer Science and Engineering, Punjabi University Patiala, Punjab 147002, India

[9] International Journal of Mechanical engineering and Technology (IJMET) Volume 8, Issue 4, April 2017, pp. 408–412 Article ID: IJMET_08_04_043 ISSN Print: 0976-6340 and ISSN Online: 0976- 6359, BIG DATA SECURITY AND PRIVACY: A SHORT REVIEW Dr. Md. Tabrez Quasim Department of Computer Science & Information System, University of Bisha, Saudi Arabia Mohammad. Meraj College of Computer & Information Sciences, King Saud University, Saudi Arabia

[10] Comprehensive Survey on Big Data Privacy Protection MOHAMMED BINJUBEIR 1, ABDULGHANI ALI AHMED 1, (Senior Member, IEEE), MOHD ARFIAN BIN ISMAIL 1, ALI SAFAA SADIQ 2,3, (Senior Member, IEEE), AND MUHAMMAD KHURRAM KHAN 4, (Senior Member, IEEE) 1Faculty of Computer Systems and Software Engineering, University Malaysia Pahang, Kuantan 26300, Malaysia 2Wolverhampton Cyber Research Institute, School of Mathematics and Computer Science, University of Wolverhampton, Wolverhampton WV1 LY, U.K. 3Center of Altricial Intelligence Research and Optimization, Torrens University, Brisbane, QLD 4006, Australia 4Center of Excellence in Information Assurance, King Saud University, Riyadh 12372, Saudi Arabia Corresponding author: Muhammad Khurram Khan (mkhurram@ksu.edu.sa) This work was supported by the Faculty of Computer System and Software Engineering, University Malaysia Pahang under the internal grant No. RDU190311 and Fundamental Research Grant Scheme (FRGS) with Vot No. RDU190113 and in part by the Researchers Supporting Project under RSP-2019/12, King Saud University, Riyadh, Saudi Arabia.

[11] Multimed Tools Appl (2018) 77:9203–9208 DOI 10.1007/s11042-017-5301-x, Advances in Security and Privacy of Multimedia Big Data in Mobile and Cloud Computing B. B. Gupta1 & Shingo Yamaguchi 2 & Dharma P. Agrawal 3 Published online: 13 November 2017 # Springer Science Business Media, LLC 2017

[12] 2018 18th IEEE International Conference on Communication Technology, A Review of Big Data Security and Privacy Protection Technology, Denglong Lv, Shibing Zhu, Huazheng Xu College of Space Information Space Engineering University Beijing, China e-mail: ldlboss519519@163.com, Ran Liu Hebei Meteorological Bureau Hebei, China



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY