The Critical Role Played by Big Data Management in Effectively Addressing the Security and Overall Privacy Concerns Through Correlation Analysis

Shahanawaj Ahamad[1]
[1]College of Computer Science and Engineering, University of Hail, Saudi Arabia
drshahwj@gmail.com

Amit Roshan[2*]
[2]Scholar, Information Technology, Dr Ram manohar Lohiya Avadh University, Ayodhya, UP
roshan.amit03@gmail.com

Melanie Lourens[3]
[3]Department of Human Resources Management, Durban University of Technology
Melaniel@dut.ac.za

Vineet Shekher[4]
[4]Associate Professor, Electrical Engineering Department, BIT Sindri, Dhanbad
vineet.ee@bitsindri.ac.in

Kapil Joshi[5]
[5]Assistant Professor, UIT, Uttaranchal University
kapilengg0509@gmail.com
https://orcid.org/0000-0003-1097-8347

Joel Alanya-Beltran[6]
[6]Professor, Electronic Department, Universidad Tecnológica del Perú
C18121@utp.edu.pe
https://orcid.org/0000-0002-8058-6229

2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE) | 978-1-6654-3789-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICACITE53722.2022.9823900

***Abstract*—Due to the widespread application of big data in a wide range of fields, there has been a significant rise in a diverse range of data assets, and numerous data analysis technologies, such as standardized data mining or quantitative analytical techniques, are speeding up the constant enlargement of the big data industry. The fact that information from several different resources has a life cycle from acquisition to disposal is a key property of big data, as is the fact that new information may be obtained via analysis, combining, and exploitation of existing data. Yet, information security & dependability challenges arise at every stage of the life cycle, rendering the protection of individually identifying data a vital goal throughout the whole process. It is possible to evaluate user behavior using different big data analytics techniques, and the collection of this information results in the violation of personal privacy. This study analyses dangers and security challenges that arise during the administration of large amounts of data by verifying the existing standards produced by international standardization bodies and doing an analysis of relevant research.**

Keywords—Big data, Security, Privacy, Information, Management, Analytical, Internet of Things (IOT), Application

I. INTRODUCTION

The term "Big Data" refers to the management of databases that are large in comparison to the capabilities of frequently used development solutions to gather, handle, or analyze that quantity of information in a timely manner. Every two years, it is anticipated that the volumes of information to be evaluated would double (IDC, 2012). All of this information is often unstructured and comes from a variety of sources. Large datasets are being generated by the so "Internet of Things" (IoT), which is generating big quantities of information that must be converted into helpful data. As the devices which are connected to the so-called "Internet of Things" (IoT) continues to grow to previously unimaginable levels, big data is becoming increasingly popular. Furthermore, it is becoming increasingly common to purchase on-demand extra computational storage and processing through public cloud suppliers in order to undertake expensive data-parallel processing operations. As a result, the volume, diversity, and usually described of the system architecture required to enable Big Data technologies might possibly exacerbate privacy and security concerns. The proliferation of Big Data, facilitated by the use of cloud services, has rendered conventional security solutions designed for private computer architectures.

Security operations must be able to operate across a heterogeneous composition of varied hardware, software platforms, and network domains, which can only be accomplished via the use of Big Data. The abstractions capacity of Software-Defined Networking (SDN) appears to be a very crucial property in this riddle computer system, as it may allow the efficient distribution of Big Data key control on top of the heterogeneous infrastructure. Since it isolates the controlling (higher) planes from the actual system architecture that is being monitored and regulated, SDN adds abstractions into the network. In compared to previous networks, where executives (if permitted to do so by hardware vendors) must formalize capabilities in view of reduced system configurations, network virtualization allows operators to start writing elevated monitoring program that define the behavior of an overall infrastructure. System (re)configuration, system development, and the application of security policies may all be simplified with the help of SDN. By implementing a layer of managers and/or using multiple controllers, it is possible to reduce the robustness disadvantage of a centrally managed SDN solution, at the very least for most critical system operates that must be regulated.

II. OBJECTIVE

The research aimed to fulfill the following objectives:

1. To study big data management and best practice for big data management
2. To study the most difficult difficulties in big data administration
3. To study the equipment and capacities for big data management
4. To study the challenges of big data in the context of data privacy and security

III. METHODOLOGY

Numerous data analysis methods, including such big data or statistical analysis, are accelerating up the big data company growth. Big data is collected, merged, and used to get new knowledge. Unfortunately, challenges with data security and dependability remain at each and every stage of

life cycle, putting data security a concern. It is possible to analyze customer behaviors utilizing big data technologies, leading in a privacy violation. These research and existing worldwide guidelines analyze the privacy and security challenges involved with huge data handling.

IV. BIG DATA MANAGEMENT

Big data management is the process of organizing, administering, and governing vast amounts of both structured and unstructured information in a controlled environment. The purpose of big data management is to guarantee that business analytics and big data analytics programmers have access to high-quality data while also keeping data costs low for users. The use of big data management strategies is becoming more common among corporations, governmental organizations, and other groups to help them deal with fast expanding pools of information, which generally consists of many tens of terabytes [4].

Rather of relying just on conventional database systems and conventional data center platforms, many big data ecosystems contain capabilities that are well-suited for collecting and analyzing non transactional kinds of information. The growing emphasis on gathering and analyzing large amounts of data is influencing the development of new data bases and designs, which often integrate large data alongside big data systems [5].

Big data management requires organisations to choose which data must be retained for regulatory purposes, which data may be thrown of, and which information should be studied in order to enhance existing business processes or create a market edge. This method requires meticulous data categorization in order for reduced amounts of data to be evaluated promptly and profitably at the end [6].

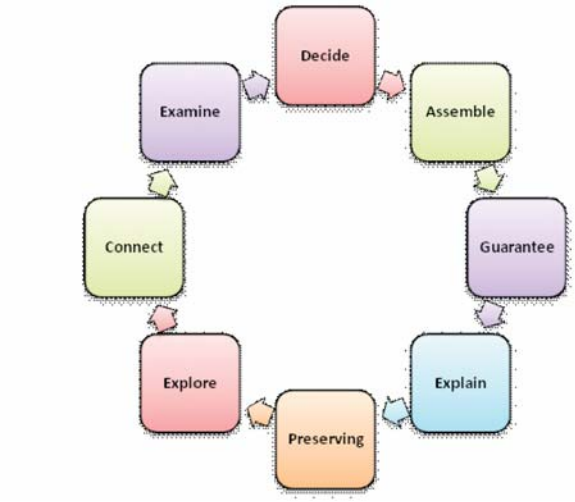

Fig. 1. Big Data Management

V. BEST PRACTICES FOR BIG DATA MANAGEMENT

Big data processing, when done correctly, lays the groundwork for promising effects programmers that may aid in improved corporate judgment & long-term planning in businesses [7]. In table 1. Here is a list of recommended practices to use in big data initiatives to get them off on the right foot:

TABLE I. BEST PRACTICES FOR DATA MANAGEMENT

1	**Create a clear plan and road map ahead of time.**	Organisations must begin by developing a big data strategy plan that specifies business objectives, analyses data needs, and maps out application and system implementations. An evaluation of data management procedures and capabilities should also be included in the plan is to determine any gaps that need to be addressed.
2	**Maintain your concentrate on the company's objectives and demands.**	When it comes to big data settings, data management organizations must collaborate directly with data professionals, other analytics, or enterprise customers to ensure that they satisfy the information demands of the company and allow for more information choices [8].
3	**Be adaptable in your data management.**	For machine learning, prescriptive modeling, as well as other forms of big data analytics apps, data analysts often want the ability to personalize how they alter information, and in certain situations, they require the ability to study whole amounts of primary data. An iterative strategy to data processing and organization is thus required [9].
4	**Ensure that accessibility & administrative rules are rigorously enforced.**	Although controlling massive data is difficult, it is necessary, coupled with strict user access rules and information security safeguards. This is largely to assist enterprises in complying with information privacy rules that control the gathering and the use of private information, but very well information may also result in increased and much more reliable statistics.
5	**Disconnecting data silos should be removed.**	A big data design should be developed without soloed systems to reduce data connection issues and guarantee that appropriate information is available for analysis. It also allows for the connection of current information silos as multiple sources, allowing them to be integrated with other huge datasets.
6	**Establishing a robust design and putting it into action**	An addition big data design involves many levels of technologies and tools that enable information management operations such as data intake, analysis, and preservation, as well as quality of the data, connectivity, and preparations [10].

VI. THE MOST DIFFICULT DIFFICULTIES IN BIG DATA ADMINISTRATION

Big data is typically complicated; in additional to its volume and diversity, it frequently contains broadcasting information as well as other forms of information that are generated and modified at a rapid pace. As a consequence, massive data analysis and management are difficult jobs [11]. In table 2. We show the following are the most significant obstacles for data management organizations when it comes to big data deployments:

TABLE II. CRITICAL DIFFICULTIES

1	**Resolving issues with data integrity**	Data from diverse sources databases that may not have been input or structured properly is common in big data settings, and it is important to understand how to deal with this kind of information. Teams have a difficult task in data quality management because they must discover and correct data mistakes, variations, duplicated entry, or other problems in data sets [12].

2	**Organizing data in preparation for analytics applications**	Data preparation for advanced analytics may be a time-consuming procedure, which is made even more difficult by the presence of large amounts of data. Raw data collections are often need to be aggregated, filtered, structured, and verified on the fly in order to be used in certain applications. The dispersed nature of big data systems makes it more difficult to collect the information that is necessary.
3	**Big data collections that are governed**	Without adequate data administration control, data from many sources may not be harmonized, and sensitive information may be acquired and utilized in an inappropriate manner. However, given of the unstructured & semi-structured data that big data environments include, as well as the widespread integration of external databases [13], managing big data environments presents additional issues.
4	**Ensure that large data solutions can scale up and down as required**	Big data workloads need a significant investment in storage and distribution resources. This may put a burden on the efficiency of big data systems, especially if they are not intended to handle the amount of data that is being processed. It is, however, a delicate balancing act: Businesses incur excessive expenditures as a result of the deployment of systems with excess capacity.
5	**Bringing together a variety of data types**	The information integrating procedure with big data is difficult by the requirement to gather combine data from diverse sources for analytical purposes, which is comparable to the issue of controlling data quality. Furthermore, because of the diversity and speed with which big data is processed, standard extract, transform, and load (ETL) integration methodologies are often unsuitable for dealing with it.
6	**Being able to deal with vast volumes of information**	Big data sets aren't had to be enormous, but they are often huge, and in many situations, they are colossal in scope and size. Aside from that, data is typically dispersed over a variety of computing systems and data storage repositories. The sheer quantity of information that is often included makes it challenging to handle all of the information in an efficient manner [14].

VII. Equipment and Capacities for Big Data Management

There are a plethora of technologies and applications accessible for handling large amounts of information, with the both source code and corporate variants of many of them being made accessible. It is possible to implement a variety of big data management, mostly in conjunction with one another, including distributed system structures such as Hadoop ecosystem, streaming data engines, cloud - based storage resources, array management system, NoSQL datasets, data store and information center systems, and SQL query engines [15].

Today, big data loads are now being performed in the cloud, where enterprises may either build up their personal systems or take use of managed services offers. This allows for simpler expansion and greater versatility during installation. Cloud platform industry leaders AWS, Google, & Microsoft, as well as Cloud era, Data bricks, as well as other companies that specialize in big data technologies, are among the leading providers of big data management solutions.

Standard data management technologies are also important elements in the management of large amounts of data. This includes the data assimilation software that supports numerous assimilation methods, including such conventional ETL processes; a suitable ELT reach that fills up information into big data systems in its raw form [16]; and information connectivity software that supports multiple connectivity methods, including such traditional ETL processes. The data is stored in a format that may be converted later if necessary; and genuine ensure the consistency such as modification data gathering are used. As well as manual characterization and cleaning, data integrity solutions that automate the process are also often utilized.

Fig. 2. Big Data Equipment & Capacities

VIII. The Challenges of Big Data in the Context of Data Privacy & Security

Expanding data collection, storage, and processing volume is being generated by the expansion of devices that are both linked to the Web and connected to one another. This poses new issues in terms of confidentiality as more gadgets become linked to one another. In fact, existing security methods including such as security methods cannot be employed in the big data design, gateways or DMZs cannot be used. must be extended beyond the boundary of the company's network in order to meet the requirements of user/data movement as well as the regulations of BYOD (Bring Your Own Device) [17]. The essential issue, in light of these new situations, is: What privacy & security regulations & techniques are better appropriate for meeting the daily top Big Data security & privacy needs [18]? Big Data issues may be divided into four categories: infrastructures protection, information privacy & security of personal information, data management and consistency, and responsive security. Assess the risks associated with Big Data. There are a number of risk categories that must be evaluated. The data longevity, the data development and gathering process, and the absence of security processes are examples of these issues. In the end, the Big Data security goals are no differently from those of any other sort of data - to maintain the confidentiality, integrity, and availability of the data.

Considering that Big Data is such a significant and complicated subject, it is probably inevitable that significant security and privacy issues will surface [19]. Big Data has

certain features that have an impact on information security. These qualities include variety, volume, velocity, value, variability, and truthfulness, among others. These difficulties have a direct influence on the design of security solutions, which are needed to address all of the traits and needs mentioned above. At this time, there is no such "out of the box" monitoring system available.

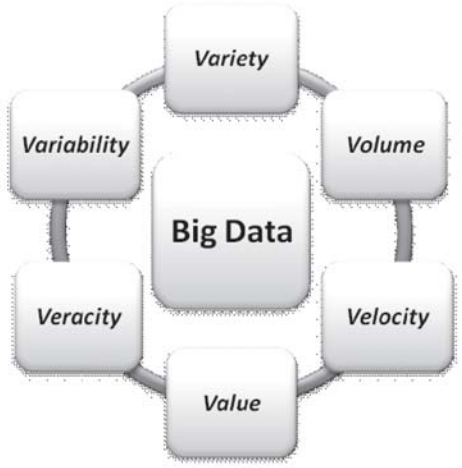

Fig. 3. Vs of Big Data

The Cloud Safe Alliance (CSA), a non-profit organization whose aim it is to encourage the adoption of best practices for applying security assurances inside Cloud Computing, has formed a Big Data Discussion Team that has concentrated on the primary difficulties to implementing secured Big Data services [20]. The CSA has divided the many security and privacy problems into four distinct parts of the Big Data environment. Facilities security, data protection, information management & consistency, and responsive security are examples of these characteristics. As per the CSA, each of these elements has the associated security challenges:

- Asset Security

1. Secure Integrated Information Processing
2. Non-Relational Data Security Best Practices

- Information Security

1. Information Extraction for Information Analysis Maintaining Information Privacy
2. Cryptographic Information Security Solutions
3. Authentication And authorization with Granularity

- Information Reliability and Administration

1. Transactions Records and Secure Information Storage
2. Inspects at the Microscopic Level Origin of Data

- Security that reacts

1. Filtration and Verification from Start to Finish
2. Real-Time Monitoring of Security Levels

Such privacy and security problems span the whole Big Data lifespan.[21]. The massive increased number of devices connected has prompted manufacturers to release a wide range of mobile apps, cloud infrastructure, and gadgets in a limited number of times in order to capitalize on this opportunity. While technology offers significant advantages and potential to Also it raises security concerns for end consumers.

HP has completed research on the market's availability IoT products and discovered that 70 percent of them had security flaws. These security breaches were attributed to privacy concerns, a lack of authorization, a lack of parking encryption, an unsecured online interface, or insufficient software protection [21]. Depending on some of these results, HP has launched an OWASP (Open Web Application Security Initiative) project called "OWASP Internet of Things," with the goal of assisting IoT manufacturers in identifying highest level of security Issues about IoT devices and how to avoid them This initiative, such as the OWASP top, uncovered the security weaknesses listed in table 3 below.

TABLE III. SECURITY WEAKNESS

1	**Inadequate Authentication /Authorization**	Inadequate Authentication/Authorization allowing the attacker to leverage a faulty password protection, crack incorrect pass codes in order to get login to sensitive settings on an IoT system.
2	**Inadequate Transportation Encrypt**	An adversary may spy on data in motion among st IoT devices or support networks due to a lack of Transport Encryption [23].
3	**Unsafe Cloud Interface**	An adversary may utilize numerous known attacks (insufficient authentication, lack of necessary encrypt, user enumeration) to obtain information or functions through the cloud portal if the public connection is insecure.
4	**Inadequate Security Configuration**	Inadequate Security Configuration: An attacker may access the data or control on the equipment owing to a lack of or inadequate reconfiguration procedures.
5	**Inadequate Physical Security**	If the IoT device is reasonably available, a hacker may exploit USB ports, SD cards, or other storage and disposal to get control to the device's operating system and possibly any data saved on the unit.

IX. CONCLUSION

Big data has various benefits and a bright prospect for development in a wide range of disciplines, but it also has a number of drawbacks and obstacles. As stated all through this article, although some significant measures are being taken to address There is a far way to go in terms of big data protection and confidentiality issues. The study would like to mention to some intriguing themes where the academic community might actively collaborate to build new Massive Security and privacy solution at the end of this section. Data production (including the big data sources - devices), information in this big data environment, academic concerns include data collection & transmission, data interpretation & analysis, and, ultimately, data usage.

REFERENCES

[1] N. Dewandre, "Big Data: From modern fears to enlightened and vigilant embrace of new beginnings", Big Data & Society, vol. 7, no. 2, p. 205395172093670, 2020. Available: 10.1177/2053951720936708.

[2] M. Mladin, "Supporting Regulatory Measures in the Context of Big Data Applications for Smart Grids", Frontiers in Big Data, vol. 4, 2021. Available: 10.3389/fdata.2021.675461.

[3] K. Benzekki, A. El Fergougui and A. Elbelrhiti Elalaoui, "Software-defined networking (SDN): a survey", Security and Communication Networks, vol. 9, no. 18, pp. 5803-5833, 2016. Available: 10.1002/sec.1737.

[4] D. Firmani, F. Leotta, F. Mandreoli and M. Mecella, "Editorial: Big Data Management in Industry 4.0", Frontiers in Big Data, vol. 4, 2021. Available: 10.3389/fdata.2021.788491.

[5] H. Wubetie, "Missing data management and statistical measurement of socio-economic status: application of big data", Journal of Big Data, vol. 4, no. 1, 2017. Available: 10.1186/s40537-017-0099-y.

[6] A. Smith, "Some Best Practices in Big Data Management", Current Trends in Biomedical Engineering & Biosciences, vol. 7, no. 3, 2017. Available: 10.19080/ctbeb.2017.07.555712.

[7] M. Day-Calder, "How to stay focused on your goals in 2018", Nursing Standard, vol. 32, no. 22, pp. 37-38, 2018. Available: 10.7748/ns.32.22.37.s31.

[8] P. Bassett, "Managing for flexible software manufacturing", Computer, vol. 31, no. 7, pp. 100-102, 1998. Available: 10.1109/2.689681.

[9] F. MENG, "Design and implement of redundant OPC data acquisition system", Journal of Computer Applications, vol. 29, no. 1, pp. 256-257, 2009. Available: 10.3724/sp.j.1087.2009.00256.

[10] "Challenges in Data Quality and Complexity of Managing Data Quality Assessment in Big Data", Regular, vol. 9, no. 3, pp. 589-593, 2020. Available: 10.35940/ijrte.c5643.099320.

[11] A. N.J., "Algorithmic Assessment of Text based Data Classification in Big Data Sets", Journal of Advanced Research in Dynamical and Control Systems, vol. 12, no. 4, pp. 1231-1234, 2020. Available: 10.5373/jardcs/v12sp4/20201598.

[12] R. Jutta, M. Karlheinz and Ď. Matej, "Linguistic Variation in the Austrian Media Corpus. Dealing with the Challenges of Large Amounts of Data", Procedia - Social and Behavioral Sciences, vol. 95, pp. 111-115, 2013. Available: 10.1016/j.sbspro.2013.10.629.

[13] S. Shams and L. Solima, "Big data management: implications of dynamic capabilities and data incubator", Management Decision, vol. 57, no. 8, pp. 2113-2123, 2019. Available: 10.1108/md-07-2018-0846.

[14] P. Obitade, "Big data analytics: a link between knowledge management capabilities and superior cyber protection", Journal of Big Data, vol. 6, no. 1, 2019. Available: 10.1186/s40537-019-0229-9.

[15] M. Kantarcioglu and E. Ferrari, "Research Challenges at the Intersection of Big Data, Security and Privacy", Frontiers in Big Data, vol. 2, 2019. Available: 10.3389/fdata.2019.00001.

[16] R. Bao, Z. Chen and M. Obaidat, "Challenges and techniques in Big data security and privacy: A review", Security and Privacy, vol. 1, no. 4, p. e13, 2018. Available: 10.1002/spy2.13.

[17] K. Michael and K. Miller, "Big Data: New Opportunities and New Challenges [Guest editors' introduction]", Computer, vol. 46, no. 6, pp. 22-24, 2013. Available: 10.1109/mc.2013.196.

[18] "Cloud Computing Security, Privacy and Forensics: Issues and Challenges Ahead", International Journal of Recent Trends in Engineering and Research, vol. 4, no. 3, pp. 10-13, 2018. Available: 10.23883/ijrter.2018.4083.xwpna.

[19] "Gartner warns of big data security problems", Network Security, vol. 2014, no. 6, p. 20, 2014. Available: 10.1016/s1353-4858(14)70062-5.

[20] P. Mor, H. Sharma and A. Sodhi, "Smart Transport Management System Implemented Using Two-Way Encryption Mechanism in RFID", International Journal of Future Computer and Communication, vol. 3, no. 4, pp. 242-246, 2014. Available: 10.7763/ijfcc.2014.v3.304.

[21] V. Panwar, D. Kumar Sharma, K. Pradeep Kumar, A. Jain and C. Thakar, "Experimental investigations and optimization of surface roughness in turning of en 36 alloy steel using response surface methodology and genetic algorithm", Materials Today: Proceedings, 2021. Available: 10.1016/j.matpr.2021.03.642.