SECURITY AND PRIVACY ISSUES RELATED TO BIG DATA AND BIG DATA ANALYTICS

***SAMAILA AJI; & **SANUSI MUHAMMED ALIYU**
**Department of Mathematics and Computer Science, School of Sciences, College of Education Billiri, Gombe State, Nigeria. **Department of Health Information Management College of Health Sciences and Technology Kaltungo, Gombe State, Nigeria.*

ABSTRACT
Big data may be defined as a term used for very large data sets that have a more varied and sophisticated structure. These characteristics usually correlate with additional difficulties in storing, analyzing, and applying further procedures or extracting results. Big data analytics is the term won't to describe the method of researching massive amounts of complex data to reveal hidden patterns or identify secret correlations. However, there's a clear contradiction between the safety and privacy of massive data and therefore the widespread use of big data. This paper focuses on the categories of privacy and security of big data and big data analytics such as; Differential Privacy, Identity based anonymization, hiding a needle during a haystack, privacy-preserving big data publishing. Existing issues or challenges of privacy and security associated with big data and big data analytics domain were identified, possible solutions to those issues were also provided which include; Data encryption, Data trustworthiness, Securing IT infrastructures, and Monitoring data usage within the big data ecosystem. Finally, the paper recommends that further research is often conducted to create more robust and complicated security mechanisms to deal with the impending privacy and security issues in big data and therefore the big data analytics ecosystem.

INTRODUCTION
Big data is a collective term pertaining to data that is so large and sophisticated that it exceeds the processing capability of conventional data management systems and software techniques (Yayeh M. Y and Berie T. G., 2016). Big data specifically refers to data sets that are so large or complex that traditional

processing applications aren't sufficient. It's the massive volume of data; both structured and unstructured that inundates a business on a day-to-day basis. Thanks to recent technological development, the quantity of knowledge generated by the web, social networking sites, sensor networks, healthcare applications, and lots of other companies, is drastically increasing day by day (Jain P, Gyanchandani M, and Kharen, 2016).

Characteristics of Big Data and Big Data Analytics

Big Data has specific characteristics that affect information security, these include variety, volume, velocity, value, and veracity. These big data characteristics are popularly referred to as the five Vs of big data (Al-shomrani A, Eassa F, Jambi K., 2018) and (Bashari B R, Akbarzadehb N, Ataeic P, and Khakbizd Y.,2016).

Figure 1: 5Vs of Big Data [24]

Adapted from Babak Bashari Rada, Nafisseh Akbarzadehb, PouyaAtaeic, and YasamanKhakbizd (Security and Privacy Challenges in Big Data Era) P. 438.

A. Variety: Variety is another fascinating aspect of big data, meaning that this data is available as structured, unstructured, or semi-structured form, making it extremely challenging for placement during an electronic database, especially since the generated data is in unstructured form, makes it crucial for data analysts to understand the category to which big data belongs.

B. Volume: the potential of processing large amounts of knowledge may be a critical aspect of big data especially since volume is one among of the most important challenges of conventional Information Technology structures during which companies are unable to process their large amounts of archived data logs.

C. Velocity: This points to the high speed at which data is made, processed, stored, and analyzed by the electronic database additionally to the speed at which new data is generated and moved around just like the way information on social media goes viral during a matter of seconds or the hundred hours of video content uploaded to YouTube daily.

D. Value: This refers to the complex, advanced, predictive, business analysis and insights related to the big data sets. Although there are great potential values within the usage of big data unless there's a return on investment (value generated) for the company; it might be very costly (and useless) to implement Information Technology infrastructure systems to store big data.

E. Veracity: When handling big data, there is always the likelihood of receiving cloudy data (which isn't 100% correct). The information quality and accuracy of the study largely depend upon the veracity of the information source.

Definitions of Privacy and Security

Privacy is the privilege to possess some control over how personal information is collected and used. Information privacy is the capacity of a private or group to prevent information about themselves from becoming known to people aside from those they provide the knowledge to. One serious user privacy issue is that the identification of private information during transmission over the web (Koo J, Kang G, and Kim Y, 2020).

Privacy may be a key lens through which many new technologies, and above all new surveillance technologies, are critiqued (Friedewald M, Finn R, Wright D., 2013).

Data privacy may be a comprehensive process integrating the protection of knowledge at the info generation, storage, and processing stages throughout the large data lifecycle (Al-shomrani A, Eassa F, Jambi K., 2018).

Also consistent with Al-shomrani A, Eassa F, Jambi K., (2018), Privacy is the claim of people, groups, or institutions to work out for themselves when, how, and to what extent information about them is communicated to others.

While Security is that the practice of defending information and data assets through the utilization of technology, processes, and training from unauthorized access, disclosure, disruption, modification, inspection, recording, and destruction (Koo J, Kang G, and Kim Y, 2020).

Privacy vs. security: Data privacy is concentrated on the utilization and governance of individual data, things like putting or add your own word fixing policies in place to make sure that consumers' personal information is being collected, shared, and utilized inappropriate ways. Security concentrates more on protecting data from malicious attacks and therefore the misuse of stolen data for profit. While security is key for shielding data, it's not sufficient for addressing privacy (Jain P, Gyanchandani M, and Khare n, 2016).

CATEGORIES OF PRIVACY AND SECURITY

The advent of computers required the adoption of specific means to safeguard information from unauthorized parties. the convenience of collating and processing personal data mirrors the problem to seek out legal means to ensure effective privacy on internet (Acquisti A, n.d).

However, recent technological advances have meant that they're not capable capture the range of potential privacy issues which must be addressed. Specifically, technologies like whole-body imaging scanners, RFID-enabled travel documents, unmanned aerial vehicles, second-generation DNA sequencing technologies, human enhancement technologies, and second-generation biometrics raise additional privacy issues to necessitate the expansion of obtainable security techniques to supply more sophisticated and robust means of building a reliable privacy and security measures (Friedewald M, Finn R, Wright D., 2013). Additionally, the privacy of private data refers to data protection issues. The close coupling that has occurred between computing and communications, particularly since the 1980s, these two concepts are mentioned as information privacy.

Privacy and security in terms of massive data is a crucial issue. The big data security model is not suggested within the event of complex applications, thanks to which it gets disabled by default. However, the absence of privacy and security to enrich the knowledge within the big data ecosystem makes it susceptible to the knowledge stored in big data. As such, there's a requirement to pay maximum attention to privacy and security issues. Big data analytics in

various organizations concentrate to a considerable portion of them decide to not utilize these services due to the absence of ordinary security and privacy protection tools. the subsequent are the identified categories of privacy and security in big data and large data analytics as illustrated by Jain P, Gyanchandani M, and Khare N., (2016) in their published paper titled "Big data privacy: a technological perspective and review".

1. Differential Privacy: May be a technology that gives researchers and database analysts a facility to get useful information from the databases that contain personal information of individuals without revealing the private identities of the individuals. this is often done by introducing a minimum distraction to the knowledge provided by the database system. The distraction introduced is large enough in order that they protect the privacy and at an equivalent time sufficiently small in order that the knowledge provided to analysts remains useful.

2. Identity-based anonymization: This system encountered issues when successfully combined anonymization, privacy protection, and large data techniques to research usage data while protecting the identities of users. Intel's human factors engineering team wanted to use website access logs and large data tools to reinforce the convenience of Intel's heavily used internal web portal. to guard Intel employees' privacy, they were required to get rid of Personally-Identifying Information (PII) from the portal's usage log repository but during a way that didn't influence the use of massive data tools to try to analyze or the power to re-identify a log entry to research unusual behavior.

3. Hiding a needle during a haystack: Existing privacy-preserving association rule algorithms modify original transaction data through noise addition. However, this work maintained the first transaction within the noised transaction because the goal is to stop data utility deterioration while prevention privacy violation. Therefore, the likelihood that an untrusted cloud service provider infers the important frequent item set remains within the method. Despite the danger of association rule leakage, provide enough privacy protection because this privacy-preserving algorithm is predicated on the "hiding a needle during a haystack" concept.

4. Privacy-preserving big data publishing. The publication and dissemination of data are crucial components in commercial, academic, and medical applications with an increasing number of open platforms, like social networks and mobile

devices from which data could be gathered, the quantity of such data have also increased over time. Privacy-preserving models broadly fall under two different settings, which are mentioned as input and output privacy. In input privacy, the first concern is publishing anonymized data with models like k-anonymity and l-diversity.

5. Fast anonymization of massive data streams: Big data related to timestamps is named an enormous data stream. Sensor data, call centre records, click streams, and health- care data are samples of big data streams. Quality of services (QoS) parameters like end-to-end delay, accuracy, and real-time operation are some constraints of massive data stream processing. the foremost pre-requirement of massive data stream mining in applications like health care is privacy-preserving. one among the common approaches to anonymize static data is k-anonymity.

ISSUES OF PRIVACY AND SECURITY IN BIG DATA AND BIG DATA ANALYTICS DOMAIN

The rapid climb of worldwide data by both individuals and corporations is partially led to the unexpected rise of unstructured data like photos, videos, and usually what social media has introduced to us and is predicted to continue by a dramatic increase rate of 4300% in annual data generation by 2020 making data production 44 times greater within the year 2020 as compared to 2009 (Bashari B. R and Ataei P,2016).

With the proliferation of devices connected to the web and connected, the quantity of knowledge collected, stored, and processed is increasing a day, which also brings new challenges in terms of data security. In fact, the currently used security mechanisms like firewalls can't be utilized in the big data infrastructure because the safety mechanisms should be stretched of the perimeter of the organization's network to satisfy the user/data mobility requirements and therefore the policies of BYOD (Bring Your Own Device). Considering these new scenarios, the pertinent question is what security and privacy policies and technologies are more capable to fulfill the present top Big Data privacy and security demands. These challenges could also be organized into four Big Data aspects like infrastructure security (e.g. secure distributed computations using MapReduce), data privacy (e.g. data processing that preserves privacy/granular access), data management (e.g. secure data

provenance and storage) and, integrity and reactive security (e.g. real-time monitoring of anomalies and attacks) (Moura J and Serrao C,2015).

There are several problems with privacy and security associated with big data. during this research, five (5) privacy and security issues associated with big data and big data analytics are outlined as mentioned by Bashari B R, Akbarzadehb N, Ataeic P and Khakbiz Y., (2016) in their journal titled "Security and privacy challenges in Big Data Era", these issues are; i. Hadoop ii. Infrastructure (Cloud) iii. Monitoring and auditing iv. Key management v. Data security.

According to Moura j and Serrao C., (2015), identify four security and privacy issues that cover the entire aspect of massive data, these issues include:

1. Infrastructure Security (Secure distributed processing of knowledge and security best actions for non-relational databases).
2. Data Privacy (Data analysis through data processing preserving, cryptographic solutions for data security and granular access control).
3. Data Management and Integrity (Secure data storage and transaction logs, granular audits, and data provenance)
4. Reactive Security (end-to-end filtering and validation and supervising the safety level in real-time). Moura, J, and Serrao, C (2015) also revealed that HP researched on the web of Things (IoT) solutions supported the market-available, which concluded that 70% of the problems are security issues like; privacy issues, insufficient authorization issues, lack of transport encryption, insecure web interface and inadequate software protection.

Zhang, D, (2018), also identified the subsequent because the big data security and privacy challenges; privacy risks, big data privacy protection technology is lacking, big data credibility must be confirmed and threat to data security.

Cuzzocrea A, (2014), a printed paper titled "Privacy and Security of massive Data: Current Challenges and Future Research Perspectives "stated a number of the present issues on privacy and security for giant data, these issues include but not limited to;

i. Privacy-Preserving Social Network Mining.
ii. Security problems with (Big) Outsourced Databases
iii. Privacy-preserving Big Data Analytics
iv. Big Data Exchange: Security Aspects
v. Privacy-Preserving Big Graph Analysis and Mining
vi. Querying Cloud-Enabled DBMS

POSSIBLE SOLUTIONS TO ADDRESS BIG DATA SECURITY AND PRIVACY CHALLENGES

Due to its characteristics, Big Data projects got to take a holistic vision of security. there is no single miraculous solution to unravel the identified Big Data security and privacy challenges and traditional security solutions, which are mainly dedicated to protecting small amounts of static data, are not capable the many nuts and bolts imposed by Big Data services (Moura, J, and Serrao, C., 2015).

Below are a number of the possible solutions to the safety and privacy issues associated with big data and big data analytics:

1. Data Trustworthiness

Big Data security and privacy are necessary to make sure data trustworthiness throughout the whole data lifecycle from data collection to usage. The new big data security solutions should extend the secure perimeter from the enterprise to the general public cloud. A trustful data derivation mechanism should even be created across domains. Besides, similar mechanisms are often won't mitigate distributed denial-of-service (DDoS) attacks launched against big data infrastructures (Moura, J, and Serrao, C., 2015).

2. Data Encryption

There is a requirement to adopt a way for the personalization feature of some big data services and their impact on user privacy. this will successfully be achieved through encoding. Encryption may be a method of remodeling understandable data (e.g. plaintext) into incomprehensible form (e.g. ciphertext). Encryption also refers to the method of converting plaintext into ciphertext through a mathematical algorithm. This suggests only those that have a group of encryption keys can change the ciphertext to plaintext, which is named decryption. ABE may be a sort of public-key cryptography that performs encryption and decryption supported an object attribute set and therefore the access structure. it's that decryption is feasible only the attribute of the ciphertext and therefore the user attribute is about to match. ABE is split into KP-ABE and CP-ABE. In KP-ABE, the conditions (e.g., policy) which will be decrypted are included within the user secret key, and CP-ABE is included within the ciphertext; it's susceptible to collusion attacks. Besides, it's widely utilized in Internet of Things (IoT) environments with many elements which will be used as attributes. While trying to form the foremost of big data, in terms

of security and privacy, it becomes mandatory that mechanisms that address legal requirements about data handling, got to be met. Secure encryption technology must be used to guard all confidential data (Koo J, Kang G, and Kim Y., 2020). The normal encryption technique can only protect static, but not data on communication systems just like the big data ecosystem. Other encryption techniques like Secure Function Evaluation (SFE), Fully Homomorphic Encryption (FHE), Functional Encryption (FE), and partition of knowledge on non-communicating data centers, can help to unravel the restrictions of traditional security techniques. (Moura, J, and Serrao, C., 2015).

3. Securing the IT Infrastructures

Apart from the supply of security to data, there is a requirement to require into consideration the safety of the IT infrastructure (physical security). This will be achieved by deploying security systems at the sting of the system's network. To support and secure processing, the longer term of the large data infrastructure should be supported by a corresponding security infrastructure that might ensure normal infrastructure operation, assets, and knowledge protection, and permit user identification/authentication and policy enforcement within the distributed multi-organizational environment (Demchenko Y, Ngo C, and Membrey P., 2013).

4. Monitoring Data Usage in Big Data Ecosystem

Monitoring, analyzing, and controlling data usage over the large data ecosystem helps to enhance the safety level and to leverage the prevailing data using security solutions (Moura, J, and Serrao, C., 2015).

RESEARCH DESIGN

This paper is a review paper, that means it reviews different scholarly assertions about privacy and security in the big data and big data analytics ecosystem. Various scholarly definitions of big data and big data analytics, characteristics of big data, categories of privacy and security in big data, issues of privacy and security in big data and possible solutions to the identified issues were discussed.

Data Collection Techniques

Data collection technique used for this paper were journals downloaded from prominent academic database such as Elsevier, Springer and Google Scholar.

Conclusion and Further Research

With the increase in big data trends and the need for heterogeneous organizations with diverse goals and data formats to establish a hand-shake among their databases to ease access to various sources of information. Sourcing information from diversified databases makes it to reduced cost and gain optimal productivity of goods and/or services. Establishing hand-shake among various databases with different data format makes it vulnerable to intruders to have access and consequently cause havoc to the organizations' success. Therefore, it is essential to identify the existing big data privacy and security issues and their possible solutions to call a halt to the emerging challenges to the shared data. In this paper, current issues in big data are identified. Possible solutions to the privacy and security issues are presented which include; data trustworthiness, data encryption, securing the IT big data infrastructures. Further research can be conducted to reveal modern security and privacy techniques to secure the increasing volume of data communication among heterogeneous computing devices such as; Internet of Things, Machine learning, Neural networks, and other sophisticated devices that may be invented to ease human activities.

REFERENCE

Acquisti, A. (2004). Privacy and security of personal information. In *Economics of Information Security* (pp. 179-186). Springer, Boston, MA.

Cuzzocrea, A. (2014, November). Privacy and security of big data: current challenges and future research perspectives. In *Proceedings of the first international workshop on privacy and secuirty of big data* (pp. 45-47).

Demchenko, Y., Ngo, C., de Laat, C., Membrey, P., & Gordijenko, D. (2013, August). Big security for big data: Addressing security challenges for the big data infrastructure. In *Workshop on secure data management* (pp. 76-94). Springer, Cham.

Jain, P., Gyanchandani, M., & Khare, N. (2016). Big data privacy: a technological perspective and review. *Journal of Big Data*, *3*(1), 1-25.

Moreno, J., Serrano, M. A., & Fernández-Medina, E. (2016). Main issues in big data security. *Future Internet*, *8*(3), 44.

Mortazavi, M., & Salah, K. (2015). Privacy and big data. In *Privacy in a Digital, Networked World* (pp. 37-55). Springer, Cham.

Moura, J., & Serrão, C. (2015). Security and privacy issues of big data. In *Handbook of research on trends and future directions in big data and web intelligence* (pp. 20-52). IGI Global.

Rad, B., Akbarzadeh, N., Ataei, P., & Khakbiz, Y. (2016). Security and Privacy Challenges in Big Data Era. *International Journal of Control Theory and Applications*, *9*(43), 437-448.

Zhang, D. (2018, October). Big data security and privacy protection. In *8th International Conference on Management and Computer Science (ICMCS 2018)* (pp. 275-278). Atlantis Press.