Review Article

The Security Challenges of Big Data Analytics: A Systematic Literature Review.

ABSTRACT

The huge amount of data generated from heterogeneous sources such as social networking sites, healthcare applications, sensor networks and many other sources are drastically increasing from time to time swiftly. Big Data is described as extremely large datasets that have grown beyond the capability to manage and analyze them with traditional database processing tools whereas big data analytics is the use of advanced analytic techniques against a very large heterogeneous datasets that include structured, semi-structured and unstructured data from different sources and in different sizes for the sake of better decision making, cost reduction, increasing operational efficiency and improved data-driven analysis. The larger the quantity of data by itself is not advantageous unless analyzed to produce valuable information. This deluge amount of data creates an operational risk in which, the risks arise from storage devices, security of tools or the technologies used to analyze the data. In this paper, we perform a systematic literature review to give comprehensive review of security challenges and risks related to big data analytics. Security mechanisms such as cryptographic and non-cryptographic techniques are used to secure big data during analytics. The security of big data at rest and in transit gets enough investigation while a few or no of research had done at securing data at processing stage. Even though a number of possible techniques were proposed for big data security though it still suffers performance issues. This article, is trying to explores security issues that used for preserving the CIA triad, non-repudiation as well as Access control in the context of big data analytics. Finally, we identify open future research directions for security of big data analytics. This paper also can serve as a good reference source for the development of modern security-preserving techniques to address various challenges of security and privacy-related scenarios.

***Key words**: Big Data Analytics, Homomorphic encryption, verifiable encryption, differential privacy.*

1. Introduction

According to Gartner report, the term big data is defined as Huge-volume, velocity and high-variety information resources that requires economical, innovative forms of information processing that enables us to get enhanced insight, decision making and process automation. In the last successive decades, the concept of big data is highly increasing possibly in almost all industry sectors, extensive amount of investment and research for developing novel techniques has been done to big data analytics [2]. Since its inception, big data brings enormous number of benefits and insights for organizations that utilize by analyzing it for better decision making, cost reduction and operational efficiency, improved data-driven analytics. Its complexity also increases through time. At the beginning of its initiation in the late 1980s there are only three V's as a characteristic (i.e. Volume, Velocity, Variety). Currently, with the increase of data sources and demands on the service the characteristics of big data increase to 17V's (volume, velocity, value, variety, veracity, validity, visualization, virility, viscosity, variability, volatility, venue, vocabulary, vagueness, verbosity, voluntariness, and versatility) [1]. This deluge of data does not bring only advantages but it comes with some potential challenge for the data users and organizations. As one of the big data characteristics (V's) is vulnerability the data in big data system is vulnerable to different kinds of ~~attack. data lo~~ss, data theft, modification of data, loss of privacy, denial of service is some of the common problems that big data systems are facing.

Currently data generated from varieties of sources such as IoT devices, streaming medias, social networks, financial sectors, government sectors and others sectors generate a deluge of data. This huge stream of data brings opportunities as well as challenges like high storage, processing and security challenges. In this paper, we perform a comprehensive review of security mechanisms used in big data analytics, limitations of those security mechanisms, techniques used to improve the performance of security mechanism and security threats in big data analytics thoroughly discussed.

As the data size increase dramatically, organizations mostly outsource their data to third party cloud companies. Data hosted in this companies may suffer a threat of unauthorized access, modification (updating or deletion) and sharing of information to third parties [3]. There are three scenarios' according to the cloud service provider's behavior. These are trusted cloud, semi-trusted cloud and untrusted cloud. Even though in the trusted cloud, there is a possibility of data breaches

that endanger the organizations safety and security. When security mechanism designed for big data analytics it should consider those aforementioned scenarios and the related vulnerabilities.

In [4],big data eco-systems are described as complex systems of networked architectures and heterogeneous devices interrelated components which can work together and is responsible for with different sizes, speeds and types of big data in the ecosystem. Authors in [5], data is increasingly viewed as a commodity and new form of currency. Though, big data appears to offer various benefits to organizations, it introduces challenges due to the unstated nature of the procedures done at the time of collection, processing, storage and use for different business firms. In each life cycle of BDA serious security challenges are raised when dealing with data acquisition, curation, processing, storage and usage [17, 23, 24].

There exist various types of security threats in big data analytics such as privacy breaches and leakage threats [6], model inversion attack and gradient inference attack [7], attribute linkage attacks [8], identity disclosure, link disclosure and content disclosure [9] and frequency attack [10] are some of the potential threats in big data analytics. In order to solve the aforementioned threats various methods are proposed by different scholars. From those cryptographic methods, access control mechanisms, perturbation and non-perturbation methods mentioned as a solution to secure big data analytics [6]. Cryptographic methods use mathematical algorithms to encrypt and decrypt data. Different cryptographic methods proposed as a solution for big data analytics security from the proposed cryptographic methods homomorphic encryption show a huge interest and popularity in big data analytics security among other security alternatives. Other cryptographic methods also used such as verifiable encryption, Secure Multiparty computation, and Functional encryption also used in Big Data Analytics. Non-cryptographic methods such as perturbation techniques (i.e. generalization, Differential Privacy and suppression), non-perturbation techniques such as l-diversity, k-anonymity, and t-closeness are other options to secure Big Data analytics.

The rest of this paper is structured as follows. Section 2 defines research questions. Section 3 introduces research methods in detail. Section 4 covers security mechanisms used in big data analytics Section 5 explains about tools for privacy-preserving in big data analytics. Section 6 is discussed about performance issues. Section 7 states about contributions , Section 8 concludes and provides future work suggestions.

2. Research Questions

RQ1: How to securely process data?

RQ2: What are the potential security threats in big data analytics?

RQ3: What are the methods to solve performance issues related to security mechanisms in big data analytics?

RQ4: How to preserve security of big data analytics in cloud computing platform?

3. Research Method

Despite the fact that, Big data analytics is very important for the sake of decision making, it is exposed to serious privacy breaches and security attack unless if we don't apply proper security measures. This paper examines, various different kinds security threats, preservation techniques and models with their limitations. In order to briefly understand and summarize the security concern of Big Data analytics we use Systematic Literature Review (SLR) as a methodology.

There are different benefits to use Systematic Literature Review as suggested by kitchenham and Charters [11] such as it is a means of evaluation and interpreting all available research resources and it presents a fair evaluation of a research topic using trustworthy, rigorous, and auditable methodology. We follow the following SLR steps:

1. **Planning the review**: in this section identification of the need for a review, specifying the research question(s) performed.
2. **Conducting the review**: in this phase identification of research, selection of primary studies, data extraction and monitoring ~~performed~~
3. **Reporting the review**: specifying dissemination mechanisms, formatting the main report conducted in this phase.

3.1. Planning the review

3.1.1. Identification of The Need for The Review

As far as the knowledge of authors concerned there is no comprehensive survey on security of Big Data Analytics. In order to fill that gap the authors propose this SLR methods.

3.1.2. Specifying The Research Question

all relevant questions are stated in research question section used to answer all questions related to security of big data analytics.

3.1.3. Development of Review Protocol

Research papers are taken from highly-regarded top journals and conferences in the field via well-established and acknowledged databases. A general approach is taken to break down the question into individual facets i.e. Study designs. Then come up with a list of keywords and their related words, abbreviations. Sophisticated search strings can then be constructed using Boolean AND's and OR's such as Homomorphic Encryption AND Big data analytics, Encryption AND Big data analytics.

i. IEEE Xplore(www.ieexplore.ieee.org/Xplore/)
ii. Elsevier ScienceDirect(www.sciencedirect.com)
iii. Google Scholar(www.scholar.google.com)

3.2. Conducting the review.

After a rigorous research authors understand there is a gap in securing users while in big data analytics. in order to fill the gap authors, conduct this research.

3.2.1. Inclusion criteria

Retrieved papers that directly linked to search criteria's which include big data analytics and security included for further investigation.

3.2.2. Exclusion criteria

Papers that only focuses on only one topic rejected such as paper that talk only big data analytics or security.

3.3. Reporting a review

After we come up with selected papers, we perform rigorous review, we write a compressive summary of security challenges of big data analytics.

4. Security Mechanisms used in Big Data Analytics

In this section, we will discuss about the security mechanisms that applied to big data analytics. In Big Data, various kinds of security mechanisms used. Generally, security mechanisms categorized in to two categories. Cryptographic and Non-cryptographic mechanism. In Figure 1 a classification of security mechanisms in detailed illustrated.

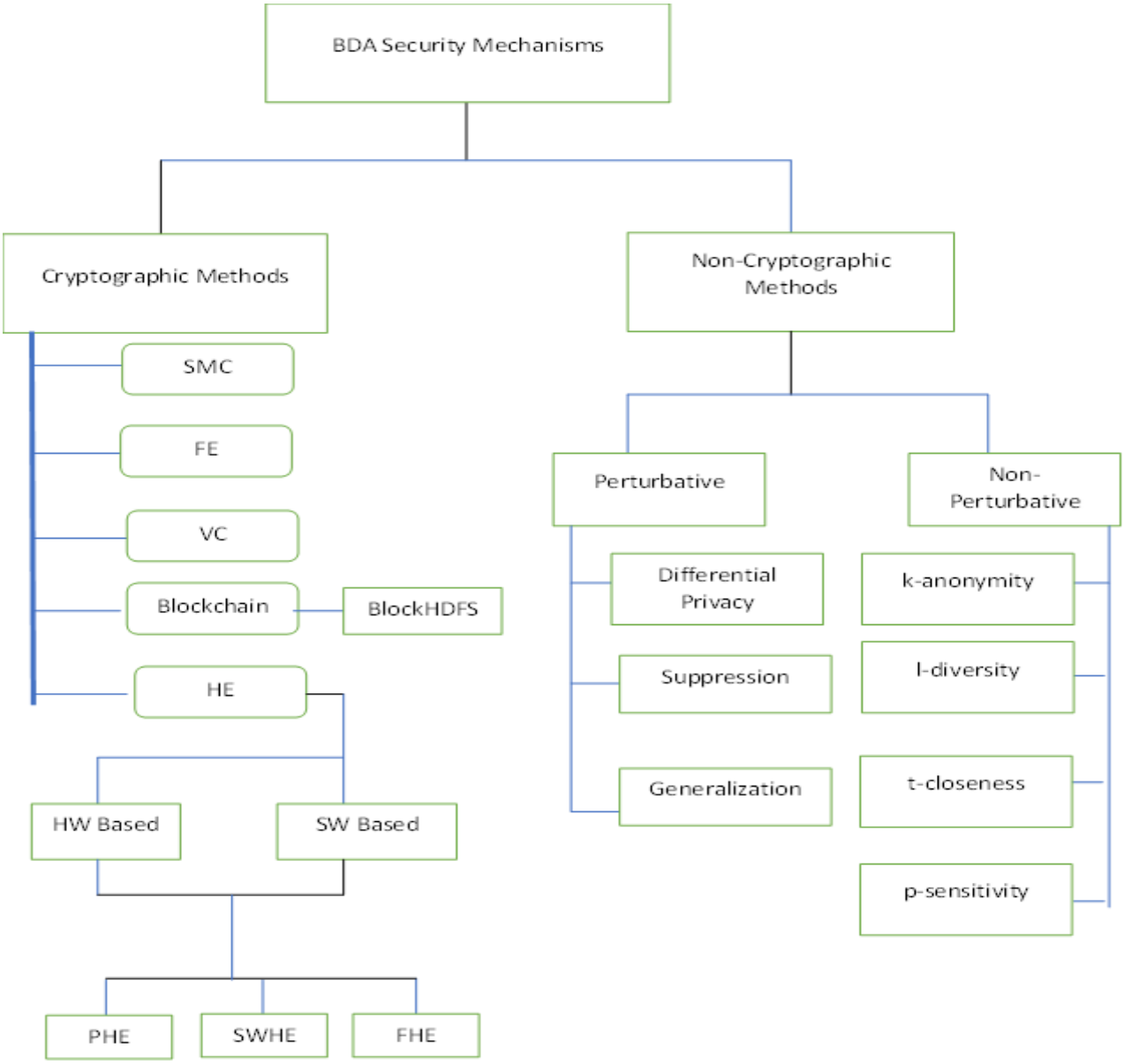

Figure 1: Big Data Analytics Security Methods

4.1. Cryptographic Security Mechanisms

Cryptographic security mechanisms use mathematical formula to make the message unreadable by unauthorized person. Some of the mentioned cryptographic mechanisms are Homomorphic encryption (HE), Verifiable Computation (VC), and Multiparty Computation (MPC) [3,11].

When we outsource our data to cloud service providers, there is a probability that cloud service providers can use it for their intentional purpose as a result in order to protect the data modification, access and sharing of data to third parties Fully Homomorphic Encryption with probabilistic

encryption is proposed [3]. Authors in [12], compare HE, VC, and MPC in three different cloud scenarios trusted, semi-trusted and untrusted environments.

Authors in [13], proposed a privacy-preserving distributed analytics framework for big data in cloud by using Fully Homomorphic Encryption (FHE) that serves as an evolving and powerful cryptosystem which can perform computations on encrypted data without decrypting it. FHE is used to protect the enormous external and internal privacy breaches and leakage threats that faced cloud computing. As mentioned in [13], the existing privacy-preserving data mining approaches have several problems such as inefficient to protect data privacy, poor performance and too much rely on a Trusted Third Party (TTP) which is considered as a security vulnerability. The authors use FHE to preserving the privacy of users' data while stored and processed in the cloud in addition to reduce the performance effect of processing encrypted data the authors use a technique called Extremely Distributed Computing (EDC). Their experimental results on FHE developed based on two parameters these are analysis performance and accuracy, for building a secure cloud-enabled application. FHE has the capability to support the operation of addition and multiplication operations simultaneously. However, the applicability of FHE schemes remain unrealistic for real-world applications because of their computational overheads. In [14], propose ECC based homomorphic encryption scheme with SMC that can improve computation and communication cost and shows the scheme has advantages in energy, communication consumption and privacy protection through the comparison experiment between ECC based homomorphic encryption and RSA& Paillier encryption algorithm. This authors also implemented ECC to the calculation of GPS data of the earthquake measurement in seismological Bureau of Fujian Province [14] to protect national secret data and it is proved that this scheme is feasible, excellent encryption effect and highly secure.

The disclosure of private data is one of the operational risks of currently existing big data platforms. The authors in [15], proposed privacy preserving scheme Stretched Homomorphic Re-Encryption Decryption (SHRED) algorithms in order to improve the security level of a scalable Big Data platform. For optimization Laplacian noise is added. The authors claim that the scheme is secured against plaintext attack by padding over deterministic cryptosystems. SHRED also ensures secure retrieval of private and public keys.

The adoption of Big Data paradigm without giving much emphasis to security considerations leads to data breaches that exposes individual and organization data to different privacy challenges. Organizations challenged to preserve the security of the data while it is collected, transmitted, stored and processed [16]. The authors argue that HE suffers limitation in performance caused by software library and used hardware type but HE can be improved for secure processing of big data sets.

In order to solve those above-mentioned limitations on HE Hardware-based approaches is now practiced to accelerate HE encryption and decryption operations. Hardware implementations such as FPGAs, ASICs, GPUs and clustering's are all the approaches that have been investigated and the result show a significant improvement on the execution of operations (i.e. encryption and decryption). Compared to normal processer computers (CPU), GPU enabled HE can enhance the processing capability of encryption 7.68x times, whereas for decryption 7.4x times improvement [16].

Authors in [7], propose a privacy-enhanced federated learning (PEFL) scheme to protect the gradients over an untrusted server and which is a highly promising for big data analytics that can trains a global model across multiple mobile devices. The scheme addresses the problem of leaking sensitive data information while we upload the vector gradient data in untrusted cloud server using Federated learning/cooperative learning mechanism by encrypting the local gradients with Paillier homomorphic cryptosystem in order to protect the local model gradients against untrusted server. The proposed solution demonstrates that it has low computation cost and high accuracy. These authors examine theoretically prove that their scheme is secure under several cryptographic hard problems and exhaustive experimental results demonstrate that PEFL (Privacy enhance federated learning) has low computation costs while reaching high accuracy in the settings of federated learning but they did not put the results in the numeric figures.

With the huge growth of data generated from different sources it is common to outsource organizations data to third party service providers. One of the most popular Big Data processing platform Hadoop only employs the Kerberos authentication protocol for controlling Big Data access. However, confidentiality and privacy are at risk. In order to solve those problems homomorphic encryption is proposed that allows the capability of computing in encrypted data without decrypting it [17]. Besides SE (Searchable encryption), PIR (Private Information

Retrieval), and MPC (Multiparty Computations) whose schemes focusing on searching, retrieving, and joint-computing in relation to encrypted data respectively. According to the number of operations supported and homomorphism properties, Homomorphic encryption permits computation on encrypted data. HE can be categorized as deterministic or probabilistic based on the probability properties of the encryption mechanisms. Unlike deterministic HE, Probabilistic HE is more preferable as it generates different cipher text for the same plaintext and secret key.

HE can also be categorized as Partial Homomorphic Encryption (PHE), Somewhat Homomorphic Encryption (SWHE) and Fully Homomorphic Encryption (FHE). PHE only supports either additive homomorphism or multiplicative homomorphism that only supports additive and multiplicative operations respectively. PHE such as Pai scheme, used in real world applications such as electronic voting protocols and biometric applications due to its performance efficiency [17] [18]. SWHE is a homomorphic cryptosystem which can perform both additive and multiplicative operations with limited number of operations. FHE supports both additive and multiplicative operations with unlimited number of operations. FHE schemes can be divided into three main categories from the perspective of algorithm design. These are lattice-based, error correcting code-based and number theoretic-based [17].

Authors in [18], proposed a merging of two technologies Network Coding and Homomorphic encryption in order to increase the robustness and throughput of wireless devices as well as to maintain data privacy. The proposed scheme provides an end-to-end data privacy, sensitive data can be stored in public clouds without worrying about data privacy and also clouds can perform advanced operations in a confidential manner.

Big data can be created with in one organization or by combining data of different organization. Based on that data processed with in the same organization is called intra big data processing. While data that belongs to different organization is called inter big data processing. As the data comes from different organization inter big data processing is more challenging. One of the challenges is security and privacy of users. To solve such kinds of problems the process of de-identification is applied by using k-anonymity, l-diversity and t-closeness to enhance privacy of users. The authors introduce privacy-preserving cosine similarity computing protocol which can efficiently calculate the cosine similarity of two vectors without disclosing the vectors to each other with lightweight multi-party random masking and polynomial aggregation techniques [19].

Organizations apply different types of cryptographic techniques to protect the confidentiality of the outsourced data. However due to the expensive computations, it limits performance in true "big data" scenarios that involve large amounts of data. The authors proposed a platform called Seabead a scheme that enables efficient analytics over large encrypted data sets which relay on symmetric encryption schemes called additively symmetric homomorphic encryption scheme (ASHE) that performs large-scale aggregations efficiently and also, they introduce a novel randomized encryption scheme called Splayed ASHE(SPLASHE) which prevent frequency attacks based on auxiliary data [10].

Large amount of data from IoT devices and other Big Data sources make data security and privacy of organizations in edge position. In order to address such issues, the authors proposed a Secure multiparty computation (Secure MPC) cryptographic technique. There are different implementations of secure MPC scheme these are secret sharing, homomorphic encryption and Yao's garbled circuits [20]. Data in Big Data environment come from various sources as a result different types of data received. In order to protect such varieties of data the authors in [9] try to describe data based on their structure and propose privacy preserving mechanisms for each data type for example cryptographic techniques (like SMC) and non-cryptographic techniques (like perturbation).

Authors in [21], present a mechanism to send data in insecure medium (i.e. insecure network) by the use of secure sum computation using homomorphic Encryption. Secure sum allows cooperating parties to compute sum of their private data without revealing their individual data to one another. For the homomorphic encryption the authors use additive homomorphic encryption technique. Secure sum allows joint parties to compute sum of their individual data without the private data being revealed to other parties. The authors propose a protocol for secure sum computation using homomorphic encryption that use symmetric key cryptography.

Four cryptographic techniques HE, VC, SMPC and Functional Encryption (FE) stated as a useful cryptographic to handling secure big-data analytics in the cloud. varieties of HE (PHE, SWHE, FHE) discussed in detail. However Homomorphic Encryption only provides confidentiality of data it does not guarantee data integrity. To provide data integrity and confidentiality a combination of Homomorphic Encryption and Verifiable Computation proposed [22]. The need to confidentiality

and privacy creates homomorphic encryption that enables to send their data in encrypted format and perform blind processing without the need to decrypt the data [23].

Authors in [6], give a clear distinction between “Privacy” and “Confidentiality”. As stated in their paper confidentiality focuses on the data itself and it is considered as “data oriented” whereas privacy includes an additional “data-owner oriented” concept. Privacy-preserving methods classified as cryptographic methods, non-perturbation and perturbation. From the cryptographic methods as stated on the paper integration of SMC and HE schemes employed the most frequently. In this paper authors objectives were: first, privacy of the input data, second, privacy of the model and thirdly, privacy of the model’s output.

In [24], authors try to show Potential threats as well as security goals of Big-Data analytics by surveying three cryptographic techniques that are applicable to secure big-data analytics in the cloud such as HE, VC, and MPC. In their paper, two types of adversaries were considered first honest-but-curious (HBC) adversary, second malicious adversary. Cloud models such as trusted, semi-trusted and untrusted cloud models considered.

In [25], the authors proposed a generic framework by combining Big Data Value Chain (BDVC) with security models. The authors suggested a comprehensive model based on three cybersecurity dimensions: These were Personal control, process and technological requirements to gain cybersecurity through BDVC. Security issues like Identity & access control, Data Availability, Data Privacy, Data Confidentiality, Data Reliability and Data Integrity aspects were considered in the proposed framework.

Authors in [26], consider three process which were big data outsourcing, big data sharing and big data management to develop a novel system architecture for securing big data in cloud called Secure Authentication and Data Sharing in Cloud (SADS-Cloud). SHA-3 hashing algorithm, MapReduce model used for splitting input file, SALSA20(128 bit, and 256 bit) encryption algorithm is applied on splited blocks of data. Compression techniques such as Clustering using Density-based Clustering of Applications with Noise (DBSCAN), Lemperl Ziv Markow Algorithm (LZMA) and Indexing using Fractal Index Tree used for data management purposes.

A Huge aggregate of data is generated from healthcare industries. To protect such kinds of data a novel design is proposed by combining machine learning and advanced security mechanisms. The

proposed system has four layers which were Data sources (structured, semi-structured and unstructured), Data storage, Security and Machine learning based application layers. Encryption techniques like AES, DES, and Blowfish were tested and blowfish showed better performance result [27].

Initially Hadoop developed for trusted environment however, researchers gradually observing security concerns on Hadoop ecosystem. The authors try to assess Hadoop framework vulnerabilities, security issues and attacks. Due to its mixture ecosystem nature different organizations IBM, MapR, Cloudera, and Hortonworks build their own security API module and packaged them to the core Hadoop ecosystems. In their paper, vulnerabilities were divided in to three categories Technology/software Vulnerabilities, Configuration/Web Interface Vulnerabilities and Network/Security Policy Vulnerabilities. Technology /software Vulnerabilities related to the technologies such as programming language used to build Hadoop framework. Configuration/Web Interface Vulnerabilities focuses on Hadoop's many default ports and IP addresses which are vulnerable to different type of attacks. Network/Security Policy Vulnerabilities since Hadoop is a mixture of different types of users and databases proper configuration policy is needed [28].

The authors in [8], discussed Big Data security and privacy in healthcare industry. Different technologies used to secure healthcare data such as Authentication (SSL and TLS), Encryption techniques, Data Masking and Access Control (RBAC & ABAC) mentioned as a mechanism to protect healthcare data in big data. In addition, De-identification techniques, HybrEx (Hybrid execution model) to protect confidentiality and privacy in cloud computing, and Identity based anonymization suggested as a security mechanism.

In order to ensure the robustness of security and to facilitate sharing in Hadoop ecosystem as well as to check the authenticity of shared data a blockchain based security mechanism i.e., BlockHDFS proposed. The proposed security mechanisms used to store the metadata in tamper-proof blocks. Authors use Hyperledger Fabric which is an authenticated user only can see the data inside Hyperledger. By logging the metadata into the blockchain BlockHDFS hardens the security of HDFS [29].

MapReduce is viable option for Big Data analytics as it supports parallel processing, high tolerance to node failure and load balancing. The authors in [30], proposed a technique to detect Intrusion

Detection by combining Machine Learning technique, Artificial Neural Network (ANN) by proposing a model MapReduce Based Intelligent Model for Intrusion Detection (MR-IMID). The proposed model captures classification of security issues based on their significance.

Authors in [31], proposed an end-to-end access control mechanism on top of existing streaming technologies like Apache Storm and Apache Kafka using ABAC. Existing technologies such as Apache Storm does not have inbuilt access control mechanism. As a result, to fill the gap, the authors proposed ABAC by introducing the concept of Secure Stream and Secure Bolts in Apache Storm. The first one is an Apache Storm stream that is glossed with Access control parameters. The user wants to process the stream first should check its privilege in Secure Bolts.

Authors in [32], explore the use of Peer-to-Peer cloud systems(P2PCS) for big data analytics as well as proposes a model that merges centralized cloud system with cooperative cloud systems. For securing the data Homomorphic Encryption, Verifiable Computation (VC) and Secure Multiparty Computation (MPC) considered as appropriate for the proposed model.

4.2. Non-cryptographic mechanisms

Cryptographic encryption techniques such as HE, suffers inefficiencies in big data processing due to many reasons such as inefficiencies in software library or hardware's [19] [16]. To alleviate this problem author in [19], propose an efficient privacy preserving scheme cosine similarity computing protocol based on lightweight multiparty random masking and polynomial aggregation technique. The protocol can calculate the cosine similarity of two vectors without disclosing the vectors to each other.

Authors in [6], develop a framework-based taxonomy of privacy-preserving mechanisms. Besides cryptographic techniques for securing privacy of big-data, others also mentioned in this paper such as perturbation (used to mask the actual data value) and anonymization (to hide the data and data owner links) are used. According to their paper, privacy-preserving protection methods are classified into cryptographic methods, non-perturbation and perturbation methods. We already discussed cryptographic mechanisms in previous section. In a non-perturbative approach, every record is composed of Identifiers (*ID*), Quasi Identifiers (*QID*), Sensitive Attributes (*SA*), and Non-Sensitive Attributes (*NSA*). a non-perturbative method preserve privacy by removing Identifiers (ID), and modifying Quasi Identifiers (QID) to QID'. Methods such as suppression and

generalization are grouped under non-perturbative category. On the other hand, perturbative methods work by disturbing original data values through noise or synthetic data. Perturbative method does not require knowledge of the distribution of records while non-perturbative methods need knowledge of the distribution of records. Methods such as k-anonymity, l-diversity and t-closeness are non-perturbative approach whereas Differential privacy (DP) is perturbative approach.

Authors in [9], identify the privacy-preserving mechanisms for different data types and propose a data taxonomy according to the structure of data. A combination of one or more data sanitizing operations such as anatomization, generalization, perturbation, suppression, permutation and slicing mentioned as a solution for maintaining privacy. In their paper, Privacy Preserving Mechanisms (PPMs) categorized based on their methodologies. Firstly, Anonymization mechanism which sanitize the data to protect sensitive information. Secondly, Obfuscation mechanisms which return gibberish reports by perturbing the original data by adding noise to the original reports. Techniques such as p-sensitive, l-diversity and t-closeness used for structured data. For numerical data Laplace mechanism mentioned whereas for categorical data exponential mechanism used. For structured and unstructured data differential privacy can be used, data such as set-valued data, genomic data and image data can be protected by differential privacy.

5. **Tools for Privacy-Preserving in Big Data Analytics**

Different tools mentioned that used for protecting privacy-preserving of Big Data Analytics in. Authors provide different privacy preserving tools that used for the secure multipart computations [20], the most common and efficient ones are discussed and listed in their paper. This tools are including libraries, implementations and frameworks examples like Failrplay, FairplayMP**,** Sharemind , VIFF , SEPIA , TASTY , SPDZ , SCAPI , Wysteria , Obliv-C , Enigma , Frigate , Chameleon , WYS , Conclav .These Tools can be for secure computation in the IoT and big data domains.

Commercial data protection tools such as IBM InfoSphere Optim Data Masking, DataGuise, Cloudera Sentry,TrustDB and CipherBase mentioned in [17] as a security tool to protect Data before sending data to third party service providers. However, issues arise when we want to

perform computation in third party cloud as the data should be decrypted that expose user privacy in addition some of them TrustDB and CipherBase require co-design of hardware and software for specific customers.

Tools such as CryptDB and Monomi used to keep the confidentiality of the data even by protecting the data the service provider using cryptographic techniques however the technique is expensive as it limits the operation in true "big data". In order to solve this the authors proposed seabed that use a novel additive symmetric homomorphic encryption scheme (ASHE) [10].

One of the enablers of big data analytics is cloud computing, cloud computing is exposed to various external and internal privacy breaches and Potential threats to big data analytics leakage threats [13]. In [16], authors discuss that there is a privacy and security threat of data leakage as organizations did not think about security when they deploy big data analytics. A threat is a potential security concern to an information system. A threat can be a person or a machine which identify a specific vulnerability and use it against an organization or individual to attack the system [28].

In a federated learning, attacks can happen when the model is uploaded to untrusted server. There is a potential threat of model inversion attack and gradient inference attack in a federated learning system [7]. Privacy protection mechanism such as k-anonymity and l-diversity are good options to protect identity disclosure [8] but failed to protect against attribute linkage attacks which can associate the victim to the sensitive attribute (SA) without having to identify his/her record. Differential privacy is a good option for attribute linkage attack however, when the data is correlated, such as time-series data applying DP is challenged to guarantee utility [6].

Authors in [9], describe potential privacy breach threats for social network data in three categories. These are identity disclosure, link disclosure and content disclosure. Identity disclosure is when a node is identified by an adversary while link disclosure is the disclosure of two nodes relationship and the content disclosure is the disclosure of privacy of nodes content.

One of the popular big data analytics framework MapReduce which does not have authentication with in Hadoop, communication between JobTracker and TaskTracker being unsecured, and also Hadoop deamons did not authenticate to each other. In their paper, authors categorize

vulnerabilities into the following three categories: Technology/Software Vulnerabilities, Configuration/Web Interface Vulnerabilities and Network/Security Policy Vulnerabilities [28].

Authors in [31],discuss the vulnerabilities of Apache Storm which does not have built-in mechanism to specify and form of access control. The authors design a fine-grained access control mechanism using Attribute Based Access Control (ABAC) and Role Based Access Control (RBAC).

In deterministic encryption each plaintext value is mapped to exactly one cipher text value. There is a possibility of frequency attack. In which an adversary can identify the most frequent one by looking at the cipher text [10].

6. Performance issues

Quality of Services (QoS) parameters like end-to-end delay, accuracy, and real time operations are some constraints of massive data stream processing. As the author in [6] , explains to measure the effectiveness and efficiency of privacy preservation techniques. Evaluation metrics can be categorized as privacy metrics and utility metrics to measure the degree of privacy of the user's satisfaction level by the system, the amount and level of protection provided by the safeguarding technology.

As cloud and network architecture is going to change, so does the big data analytical part for serving the changing demand of users. Therefore, a new scalable, efficient yet cost-effective network infrastructure must provision to this ever-changing demand of users and increase IoT devices through the use of Software Defined Networks (SDN) that can support V2X,5G and highly complicated enterprise networks. The fascinating features of SDN solves different kinds of issues that prevailing for big data applications such as data delivery, joint optimization, big data processing in the cloud data centers scientific big data architectures and scheduling issues [33].

Authors in [13], design a security mechanism to protect big data analytics using fully homomorphic encryption. In order to solve the analysis performance issue, the authors use distributed method which is called Extremely Distributed Clustering (EDC) approach which used for large scale data set clustering computations. This distributed model can speed up the data clustering process at least by a factor of two by processing different parts of a dataset and clusters in different nodes.

Authors in [7], proposed a privacy enhanced federated learning (PEFL) using a Federated learning scheme which is a machine learning technique which works in cooperative mode in order to reduce performance issues. The authors use the Distributed Selective Stochastic Gradient Descent (DSSGD) method in the local training phase to achieve encryption in distributed system and thus reduces computational cost of cryptosystems.

Authors in [18], use NC as a method to increase the robustness and throughput of wireless networks, as well as for computation in cloud in encrypted format HE can be used to maintaining data privacy in wireless networks. NC is auspicious technology to address latency issues in WSNs. Unlike traditional store-and-forward transmissions models, network coding combines receiving packets in to coded packets. One of such a scheme is Random Linear Network Coding (RLNC).

Collaborative deep learning is mentioned as one way of enabling distributed learning framework that enables sharing of input data and jointly building a deep learning model. This model also keeps the data private for each data provider [6].

For effective big data management authors in [26], implement Density-Based Clustering of Applications with Noise(DBSCAN) algorithm over cloud servers, also for indexing Fractal Index Tree which performs well in terms of insertions, deletions and searching used. For managing the size of the dataset authors suggest compression before encryption using Lempel Ziv Markow Algorithm (LZMA) compressor.

7. **Contribution**

In this paper, we present a survey reflecting the recent developments in the challenging field of security of big data analytics. Our contributions are listed as follows:

- ✓ A description of potentials security threats, that will enable to address these security threats, and mechanisms to improve performance issues
- ✓ General description and classification of the big data analytics security methods.
- ✓ Describing the challenges of privacy preservations techniques
- ✓ Show the vulnerabilities and threats that big data analytics in different phases of the Big Data Analytics life cycle.
- ✓ Discuss techniques employed to enhance the processing performance of Big Data Analytics.

- ✓ Full comprehensive survey of security mechanisms.

8. **Conclusion and future Directions**

Big data analytics is the most significant area which offers many potential benefits, innovations and have a promising future. It is a remarkable domain with a promising future, if approached correctly. The difficulty with big data comes mainly from its size, which requires proper storage, management, integration, processing, and analysis. Although Big data analytics is useful for informed decision-making purpose, it will lead to serious security challenges concerns. Hence preserving the security of big data analytics became very important. this paper, comprises of discussion on recent issues, direction, chance, and challenges of big data analytics, security concerns aimed to discuss a variety of issues in more detail.

This study identified a number of challenges from the standpoint of privacy and security in Big data analytics. The outcomes of this work have been described in two steps. Firstly, it discusses the current research focus and headings as documented in the chosen highly-regarded top journals and conferences in the field via well-established and acknowledged databases. Secondly, it argues about the data that was scrutinize and deduced from the selected literature. The study enabled us to explore the existing research in Big Data Analytics Security challenges using the well-known SLR approaches. From the primary studies, we reviewed more than 150 research papers collected from the well-known databases. In the last seven years from (2014–2021), using inclusion and exclusion criterion primary studies with the inclusion and exclusion criteria's. Moreover, there are many open issues that are still not well studied and need to be investigated by future research efforts such as combining both topics of big data with Software Defined Network (SDN).

As the declaration of GDPR stated that users have the right to manage and control their own data based on their privacy concern only and without their consent no bay shall can access their data. As a result, the combination of BlockMatrix with big data analytics needs to be investigated for the future.

Declarations

1. **Ethics approval and consent to participate**

   Not Applicable

2. **Consent for publication**

   Author hereby grants and assigns to **Springer Nature** (hereinafter called **Publisher**) the exclusive, sole, permanent, world-wide, transferable, sub-licensable and unlimited right to reproduce, publish, distribute, transmit, make available or otherwise communicate to the public, translate, publicly perform, archive, store, lease or lend and sell the Contribution or parts thereof individually or together with other works in any language, in all revisions and versions (including soft cover, book club and collected editions, advance printing, reprints or print to order, microfilm editions, audiograms and ideogram's), in all forms and media of expression including in electronic form (including offline and online use, push or pull technologies, use in databases and data networks (e.g. the Internet) for display, print and storing on any and all stationary or portable end-user devices, e.g. text readers, audio, video or interactive devices, and for use in multimedia or interactive versions as well as for the display or transmission of the Contribution or parts thereof in data networks or search engines, and posting the Contribution on social media accounts closely related to the Work), in whole, in part or in abridged form, in each case as now known or developed in the future, including the right to grant further time-limited or permanent rights. For the avoidance of doubt, all provisions of this contract apply regardless of whether the Contribution and/or the Work itself constitutes a database under applicable copyright laws or not.

3. **Availability of data and materials**

   Not Applicable

i. **Conceived and select a research area**

Both authors perform a discussion to select their research area and perform preliminary research to understand the gap in their interest area.

ii. **Collected research papers**

Both authors collect quality papers for their research work.

iii. **Sharing ideas and arrange discussion sessions**

Since both authors interested on the topic both authors actively engaged in sharing ideas arranging the discussion time.

iv. **Designing the research structure**

By referring different research papers, we select the best review paper structure.

v. **Writing a paper**

First authors write a paper based on their reading and then by authors perform hot discussion on the final writeup and write the final paper together.

List of Acronyms

Acronym	
CIA	Confidentiality, Integrity, Availability
IoT	Internet of Things
BD	Big Data
BDA	Big Data Analytics
SLR	Systematic Literature Review
HE	Homomorphic encryption
VC	Verifiable Computation
MPC	Multi-Party Computation
FHE	Fully Homomorphic Encryption
TTP	Trusted Third Party
EDC	Extremely Distributed Computing
ECC	Elliptic Curve Cryptography
RSA	Rivest–Shamir–Adleman
GPS	Global Position System
SHRED	Stretched Homomorphic Re-Encryption Decryption
FPGAs	Field Programmable Gate Arrays
ASICs	Application-Specific Integrated Circuit
GPUs	Graphics processing unit
PEFL	privacy-enhanced federated learning
SE	Searchable encryption
PIR	Private Information Retrieval
MPC	Multiparty Computations
PHE	Partial Homomorphic Encryption
SWHE	Somewhat Homomorphic Encryption
ASHE	additively symmetric homomorphic encryption
SMPC	Secure Multi-Party Computation
BDVC	Big Data Value Chain
SADS-Cloud	Secure Authentication and Data Sharing in Cloud
SHA-3	Secure Hash Algorithm 3
LZMA	Lemperl Ziv Markow Algorithm
DBSCAN	Density-based Clustering of Applications with Noise
AES	Advanced Encryption Standard
DES	Data Encryption Standard
IBM	International Business Machines
API	Application Programming Interface
IP	Internet Protocol
RBAC	Role Based Access Control
ABAC	Attribute Based Access Control
SSL	Secure Sockets Layer

TLS	Transport Layer Security
HybrEx	Hybrid execution model
BlockHDFS	Block Hdoop Distributed File System
ANN	Artificial Neural Network
MR-IMID	MapReduce Based Intelligent Model for Intrusion Detection
P2PCS	Peer-to-Peer cloud systems
ID	Identifiers
QID	Quasi Identifiers
SA	Sensitive Attributes
NSA	Non-Sensitive Attributes
DP	Differential privacy
PPMs	Privacy Preserving Mechanisms
VIFF	Virtual Ideal Functionality Framework
SEPIA	Safeguarding European Photographic Images for Access
QoS	Quality of Services
SDN	Software Defined Networks
5G	5th generation
6G	6th generation
DSSGD	Distributed Selective Stochastic Gradient Descent
NC	Network Coding
WSNs	Wireless Sensor Networks
RLNC	Random Linear Network Coding
GDPR	General Data Protection Regulation

REREFERNCES

[1] A. Panimalar.S, V. Shree.S2 and V. Kathrine.A, "The 17 V's Of Big Data," *International Research Journal of Engineering and Technology (IRJET)* , vol. 04, no. 09, 2017.

[2] S. Venkatraman and a. R. Venkatraman, " Big data security challenges and strategies," *AIMS Mathematics*, 19 July 2019.

[3] V. Kumar, R. Kumar, S. K. Pandey and M. Alam, "Fully Homomorphic Encryption Scheme with Probabilistic Encryption Based on Euler's Theorem and Application in Cloud Computing," *Advances in Intelligent System and Computing*, 2018.

[4] A. Q. G. F. K. H. a. M. I. Memoona J. Anwar1, " Secure big data ecosystem architecture challenges and solutions," *Journal on wirless commninication and Networking* , 2020/2021.

[5] H. Simo, " Big Data: Opportunities and Privacy Challenges," *Working on draft paper* .

[6] H.-Y. Tran and J. Hu, "Privacy-preserving big data analytics - A comprehensive survey," *Journal of Parallel and Distributed Computing* , vol. 134, pp. 207-218, 2019.

[7] B. C. S. Y. a. H. D. Jiale Zhang, "PEFL: A Privacy-Enhanced Federated Learning Scheme for Big Data Analytics," 2019.

[8] K. ABOUELMEHD, A. BENI-HSSANE, H. KHALOUFI and M. SAADI, "Big data security and privacy in healthcare: A Review," in *Procedia Computer Science*, 2017.

[9] M. Cunha, R. Mendes and J. P. Vilela, "A survey of privacy-preserving mechanisms for heterogeneous data types," *Computer Science Review*, vol. 41, 2021.

[10] A. Papadimitriou, R. Bhagwan, N. Chandran, R. Ramjee, A. Haeberlen, H. Singh, A. Modi and S. Badrinarayanan, "Big Data Analytics over Encrypted Datasets with Seabed," in *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation*, Savannah, GA, USA, 2016.

[11] B. Kitchenham and S. Charters, "Guidelines for performing Systematic Literature Reviews in Software Engineering," Software Engineering Group School of Computer Science and Mathematics Keele University and Department of Computer Scienc University of Durham , Durham,UK, 2007.

[12] Dr.C.Nalini and Dr.A.R.Arunachalam, "A STUDY ON PRIVACY PRESERVING TECHNIQUES IN BIG DATA ANALYTICS," *International Journal of Pure and Applied Mathematics*, vol. 116, no. 10, pp. 281-286, 2017.

[13] A. Alabdulatif, I. Khalil and X. Yi, "Towards secure big data analytic for cloud-enabled applications with fully homomorphic encryption," *Journal of Parallel and Distributed Computing*, 2019.

[14] M. Hong, "Homomorphic Encryption Scheme Based on Elliptic Curve Cryptography for Privacy Protection of Cloud Computing," 2016.

[15] V.Shoba and Dr.R.Parameswari, "A Pragmatic Approach for Privacy Preserving Healthcare Using Stretched Homomorphic Re-Encryption Decryption Algorithm," *International Journal of Advanced Science and Technology,* vol. 20, no. 7, pp. 8850-8860, 2020.

[16] R. A. Hallman, M. H. Diallo, M. A. August and C. T. Graves, "Homomorphic Encryption for Secure Computation on Big Data," in *he 3rd International Conference on Internet of Things, Big Data and Security*, San Diego, Ca, U.S.A, 2018.

[17] T. S. Fun and A. Samsudin, "A Survey of Homomorphic Encryption for Outsourced Big Data Computation," *KSII TRANSACTIONS ON INTERNET AND INFORMATION SYSTEMS,* vol. 10, no. 8, pp. 3826-3851, 2016.

[18] G. Peralta, R. G. Cid-Fuentes, J. Bilbao and P. M. Crespo, "Homomorphic Encryption and Network Coding in IoT Architectures: Advantages and Future Challenges," *electronics,* vol. 8, 2019.

[19] R. Lu, H. Zhu, X. Liu, J. K. Liu and J. Shao, "Toward Efficient and Privacy-Preserving Computing in Big Data Era," *IEEE,* 2014.

[20] M. G. Raeini and M. Nojoumian, "Privacy-Preserving Big Data Analytics: From Theory to Practice," 2019.

[21] R. Sheikh and D. K. Mishra, "Secure Sum Computation Using Homomorphic Encryption," *Springer Nature Singapore ,* 2019.

[22] H. J. S. Sidhu and M. S. Khanna, "Cloud's Transformative Involvement in Managing BIG-DATA ANALYTICS For Securing Data in Transit, Storage And Use: A Study," in *Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC)*, 2020.

[23] S. Bajpai and P. Srivastava, "A Fully Homomorphic Encryption Implementation on Cloud Computing," *International Journal of Information & Computation Technology,* vol. 8, pp. 811-816, 2014.

[24] S. Yakoubov, V. Gadepally, N. Schear, E. Shen and A. Yerukhimovich, "A Survey of Cryptographic Approaches to Securing Big-Data Analytics in the Cloud," *IEEE,* 2014.

[25] A. Z. Faroukhi, I. E. Alaouib, Y. Gahia and A. Aminea, "A Multi-Layer Big Data Value Chain Approach for Security Issues," in *The 2nd International Workshop on Emerging Networks and Communications*, Leuven, Belgiu, 2020.

[26] U. Narayanan, V. Paul and S. Joseph, "A novel system architecture for secure authentication and data sharing in cloud enabled Big Data Environment," *Journal of King Saud University – Computer and Information Sciences,* 2020.

[27] P. Kaur, M. Sharmab and M. Mittal, "Big Data and Machine Learning Based Secure Healthcare Framework," in *International Conference on Computational Intelligence and Data Science*, 2018.

[28] G. S. Bhathal and A. Singh, "Big Data: Hadoop framework vulnerabilities, security issues and attacks," *Array*, vol. 1, no. 2, 2019.

[29] V. Mothukuri, S. S. Cheerla, R. M. Parizi, Q. Zhang and K.-K. R. Choo, "BlockHDFS: Blockchain-integrated Hadoop distributed file system for secure provenance traceability," *Blockchain: Research and Applications,* vol. 2, 2021.

[30] M. Asif, S. Abbas, M. Khan, A. Ftima, M. A. Khan and S.-W. Lee, "MapReduce Based Intelligent Model for Intrusion Detection Using Machine Learning Technique," *Journal of King Saud University - Computer and Information Sciences,* 2021.

[31] S. Nambiara, S. Kalambur and D. Sitaram, "Modeling Access Control on Streaming Data in Apache Storm," in *Procedia Computer Science*, Bengaluru, India, 2020.

[32] L. A. Tawalbeh and G. Saldamli, "Reconsidering big data security and privacy in cloud and mobile cloud systems," *Journal of King Saud University – Computer and Information Sciences,* vol. 33, p. 810–819, 2021.

[33] F. R. Y. a. Q. Y. Laizhong Cui, " When Big Data Meets Software-Defined Networking : SDN for Big Data and Big Data for SDN," *IEEE Network* • , January/February 2016.

[34] C. B. C. C. K. Frederik Armknecht, " A Guide to Fully Homomorphic Encription".

[35] D. Mechkaroska and A. P.-M. a. V. Dimitrova, " SECURE BIG DATA AND IOT WITH IMPLEMENTATION OF BLOCKCHAIN," *INTERNATIONAL SCIENTIFIC JOURNAL "SECURITY & FUTURE".*