IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 23 December 2023, accepted 18 January 2024, date of publication 24 January 2024, date of current version 13 February 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3357943

RESEARCH ARTICLE

Elevating Big Data Privacy: Innovative Strategies and Challenges in Data Abundance

MOHAMED ELKAWKAGY[1], E. ELWAN[2], ALBANDARI ALSUMAYT[1], HEBA ELBEH[1], AND SUMAYH S. ALJAMEEL[3]

[1]Department of Computer Science, Applied College, Imam Abdulrahman Bin Faisal University, Dammam 314441, Saudi Arabia
[2]Department of Computer Science, Faculty of Computers and Information, Menoufia University, Shibin El Kom 32511, Egypt
[3]Saudi Aramco Cybersecurity Chair, Department of Computer Science, College of Computer Science and Information Technology, Imam Abdulrahman Bin Faisal University, Dammam 314441, Saudi Arabia

Corresponding author: Sumayh S. Aljameel (saljameel@iau.edu.sa)

This work was supported by the Saudi Aramco Cybersecurity Chair.

ABSTRACT The exponential growth of big data has ushered in transformative possibilities across various sectors, but it has also raised formidable privacy concerns. This article delves into the pressing need for enhancing big data privacy and explores innovative approaches to address this critical issue. In recent years, big data has been characterized by its immense volume, high velocity, and diverse data sources. These attributes have enabled organizations to gain unprecedented insights but have also exposed sensitive information to potential breaches. As such, ensuring the privacy of individuals and sensitive data within big data sets has emerged as a paramount concern. This article first elucidates the multifaceted nature of big data privacy, emphasizing its encompassment of privacy, confidentiality, integrity, and availability. It also acknowledges the challenges posed by existing privacy-preserving techniques, which often fall short of providing comprehensive protection for large and diverse data sets. The core focus of this article lies in presenting novel strategies and technologies designed to improve big data privacy. This article presents an innovative framework that combines advanced encryption methods, including fine-grained encryption techniques and differential privacy mechanisms specifically designed for the distinct traits of big data, like noisy techniques. To achieve this, the dataset undergoes categorization into key attributes, sensitive attributes, quasi attributes, and insensitive attributes. Subsequently, the fine-grained technique encrypts key and sensitive attributes, while the differential privacy mechanism encrypts the quasi attributes. To further substantiate the effectiveness of the proposed technique, this article references to empirical findings that demonstrate tangible improvements in big data privacy protection.

INDEX TERMS Big data, big data privacy, fine-grained encryption, perturbation technique, Hadoop, MapReduce.

I. INTRODUCTION

In the digital age, where the world is increasingly interconnected, data has emerged as a powerful currency that fuels innovation, decision-making, and transformation across industries and sectors. The term "Big Data" has become synonymous with the vast and ever-expanding volumes of information generated by our digital interactions, devices, and systems. It represents an exciting frontier of opportunities and challenges that have the potential to reshape the way we understand and navigate the world. Big Data, at its core, refers to the massive datasets that are too extensive, complex, and dynamic to be effectively processed and analyzed using traditional data management tools and methods. What sets Big Data apart is not only its sheer size but also its velocity (the speed at which data is generated and must be analyzed), variety (the diverse types and formats of data), and the value it holds when harnessed correctly [1], [2].

The authors in [3] enhanced the security of the TSZ (To, Safavi-Naini, and Zhang) homomorphic encryption scheme

The associate editor coordinating the review of this manuscript and approving it for publication was Junho Hong.

by integrating obfuscation and data encryption through the introduction of genetic algorithms. In [4], the authors provides an overview of the privacy and security concerns in IoT-cloud-based systems. Furthermore, a comparative analysis is conducted, encompassing the key privacy and security issues, definitions, categories, solutions, and architectures utilized within IoT-cloud-based systems.

Big Data is not confined to the digital realm alone. It encompasses data from a multitude of sources, including but not limited to: 1) Social Media: The posts, comments, and interactions on platforms like Facebook, Twitter, and Instagram create a rich tapestry of social data, reflecting trends, sentiments, and behaviors. 2) IoT Devices: Smart appliances, wearable technology, and sensors embedded in infrastructure generate real-time data streams, enabling the optimization of processes and the improvement of services. 3) E-commerce: Every online purchase, search query, and product review contributes to the massive datasets that online retailers and marketplaces use to personalize recommendations and predict consumer behavior. 4) Healthcare: Electronic health records, medical imaging, and wearable health trackers have revolutionized patient care, research, and disease prevention by generating vast healthcare datasets. 5) Financial Transactions: The global financial system relies on Big Data for risk assessment, fraud detection, and algorithmic trading, with billions of transactions occurring daily. 6) Scientific Research: Fields such as genomics, climate science, and high-energy physics rely on Big Data analytics to process and interpret massive datasets, driving groundbreaking discoveries.

The authors in [5] highlighted significant security and privacy concerns related to Big Data which require attention to bolster the security of data processing and computing infrastructure. The paper also provides insights into the K-Anonymity technique in Big Data privacy preservation, focusing on safeguarding individuals' identities and sensitive information from potential leaks before releasing datasets for analysis. A comprehensive survey on fine-grained access control methods for secure access to big data among different entities. Access control establishes specific permissions for accessing particular object entities. Fine-grained access control ensures the provision of a specific set of permissions to authorized individuals for accessing particular data is studied in [6] and [7].

The article [8] aims to elucidate the concept of Big Data, addressing the inherent challenges concerning privacy and security. Additionally, it explores various techniques within the realms of anonymization, encryption, and Differential Privacy specifically tailored for Big Data. This exploration involves the description, analysis, and comparative examination of these methodologies.

Some key challenges in big data include: 1) Data Volume: Big data involves extremely large datasets that can overwhelm traditional data storage and processing systems. Storing and managing these vast amounts of data require scalable and cost-effective solutions. 2) Data Velocity: Data streams into systems at high speeds, often in real-time. Handling this rapid influx of data is a challenge because it necessitates efficient data ingestion, processing, and analysis in near real-time. 3) Data Variety: Big data comes in various formats, including structured, semi-structured, and unstructured data. Integrating and making sense of diverse data types can be complex and may require advanced data preprocessing and integration techniques. 4) Data Complexity: Big data is often messy and may contain inconsistencies, missing values, and errors. Cleaning and preparing this data for analysis can be time-consuming and challenging. 5) Data Veracity: Ensuring the accuracy and trustworthiness of big data can be difficult, as it may come from various sources with varying levels of reliability. Data quality issues can lead to erroneous insights and decisions. 6) Data Privacy and Security: Managing and protecting sensitive information within big datasets is a critical concern. Ensuring data privacy and complying with relevant regulations is challenging, especially in multi-party data sharing scenarios. 7) Scalability: Scalability is crucial to handle the growth of big data over time. Systems and infrastructure must be able to scale up or down as data volumes and processing demands change. 8) Data Integration: Combining data from disparate sources can be complex. Data integration challenges include schema matching, data transformation, and ensuring data consistency across different datasets. 9) Data Analysis: Analyzing big data requires specialized tools and algorithms that can efficiently process and extract insights from massive datasets. Traditional data analysis tools may not be suitable for big data analytics. 10) Resource Allocation: Optimizing resource allocation, such as computing power and storage, is crucial to ensure cost-effectiveness and performance in big data processing. 11) Data Governance: Establishing data governance policies and practices is essential for maintaining data quality, privacy, and compliance. This involves defining roles, responsibilities, and processes related to data management. 12) Interoperability: Ensuring that various software and systems can work together seamlessly is a challenge, especially when dealing with different data formats and protocols. 13) Ethical Concerns: Ethical considerations arise when handling big data, particularly in areas like data bias, discrimination, and the responsible use of data for decision-making. Addressing these challenges in big data requires a combination of technological advancements, innovative solutions, regulatory frameworks, and a skilled workforce to harness the potential insights and benefits that big data can offer while mitigating its associated risks.

The authors in [10] provide the initial effort to integrate three fundamental Differential Privacy (DP) architectures into the operational telecommunication (telco) big data platform for data mining applications. Our findings reveal that all DP architectures exhibit a prediction accuracy loss of under 5% when employing a weaker privacy guarantee. While [11] concentrates on addressing privacy and security issues within the domain of big data, distinguishing between their individual requirements. It explores privacy methods such as HybrEx, k-anonymity, T-closeness, and L-diversity and their practical implementation in business contexts. Additionally, various privacy-preserving mechanisms have been developed to safeguard privacy at various stages of the big data lifecycle, including data generation, storage, and processing.

The different Data Perturbation methods and their respective advantages with data mining technique is merged in [12]. It concluded that geometric perturbation offers an elevated level of privacy while maintaining the utility of data for users. The authors in [13] outline the security classification, which can serve as a framework for assessing other studies related to the life cycle of big data. It pinpointed threats and security challenges that emerge throughout the life cycle of big data by validating prevailing standards established by international organizations and scrutinizing associated research. Despite the significance of security and privacy in big data, numerous standards organizations lack comprehensive coverage of requirements and technologies.

The researcher in [14] performed an empirical investigation focused on utilizing big data for intelligent security analysis methods aimed at averting cyber threats. These methods are designed to safeguard information assets by assessing and predicting risks.

The enhancement of FGEM (Fine-Grained Encryption Method) to efficiently safeguard semi-structured and unstructured data is highlighted in [15]. The improvement involves organizing such data within an enhanced tree structure, optimizing FGEM performance by accessing related nodes. Experimental results demonstrate that the enhanced FGEM exhibits reduced processing time and minimal memory usage. In this context [16], the authors utilize the Perturbation Based Encryption technique by introducing garbage values. This technique incorporates adding garbage values in two specific ways within the dataset rows. For odd-numbered rows, a noise value is multiplied as garbage, while for even-numbered rows, a garbage value is added. Subsequent to injecting these garbage values into the dataset, the AES encryption algorithm for encryption is applied. While the work in [17] focuses on addressing the issue of image perturbation for privacy preservation. The authors introduce three novel systems designed to conceal small details within images by rotating specific pixels. These models leverage two algorithms: the first employs a simulation of the firework algorithm, placing fireworks on chosen pixels and depicting sparks as rotation processes. The second system utilizes a rotation-based perturbation model employing the iterated local search algorithm (ILS) with two optimization stages. Meanwhile, the third system operates similarly to the previous one but integrates the ILS algorithm with three optimization stages.

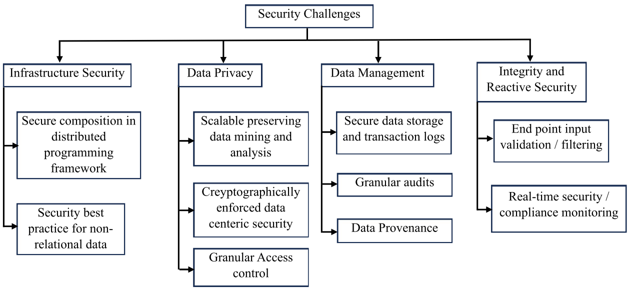
Security Challenges
Infrastructure Security
Secure composition in distributed programming framework
Security best practice for non-relational data
Data Privacy
Scalable preserving data mining and analysis
Creyptographically enforced data centeric security
Granular Access control
Data Management
Secure data storage and transaction logs
Granular audits
Data Provenance
Integrity and Reactive Security
End point input validation / filtering
Real-time security / compliance monitoring

FIGURE 1. Security and privacy challenges in the big data ecosystem.

The authors in [18] not only outlines the diverse challenges and applications of responsible and protective data privacy measures in research but also asserts that the ethical aspect of Big Data-based research should aim for a balance. This balance should consider the research's scope, utility, and subjects' privacy on one side, and the advantages and risks associated with providing access and engaging in collaborative data usage on the other. This tension reflects the significant potential of Big Data in higher education (HE) research, presenting both opportunities and difficulties for researchers. Yet, it underscores the crucial need to comprehend the extensive implications of Big Data utilization and implementation for future HE research agendas. One pivotal implication highlighted in this study is data privacy, which, as argued, can be effectively ensured through optimal information stewardship, responsible data management, and a more robust data-centric security approach.

Although there are a lot of studies discuss the big data security and privacy, none of them has yet provide robust mechanisms to effectively safeguard sensitive information within large datasets. Therefore, in this paper, we aim to enhance the fine-grained encryption technique by integrating it with the perturbation technique to bolster Big Data privacy. The proposed approach employs the map-reduce paradigm to partition the input data into numerous separate files, enabling parallel processing. It applies fine-grained encryption and perturbation techniques across distinct categories of data attributes, including key attributes, sensitive attributes, and quasi-attributes.

In the upcoming sections, we will cover various aspects of our paper. Section II includes two forks, the first one outlines the security challenges in Big Data, including an exploration of big data privacy and privacy-preserving techniques, and the second one delves into related work, while Section III introduces our proposed enhancements to fine-grained techniques for preserving Big Data privacy. Section IV discusses the results, and lastly, Section V presents the conclusion.

II. RELATED WORK

Due to the distinctive characteristics of Big Data, security and privacy concerns have significantly expanded, as highlighted by the Cloud Security Alliance (CSA). CSA's insights categorize the security challenges in Big Data into four primary domains, encompassing the entire lifecycle of Big Data, from its generation to processing, storage, transport,

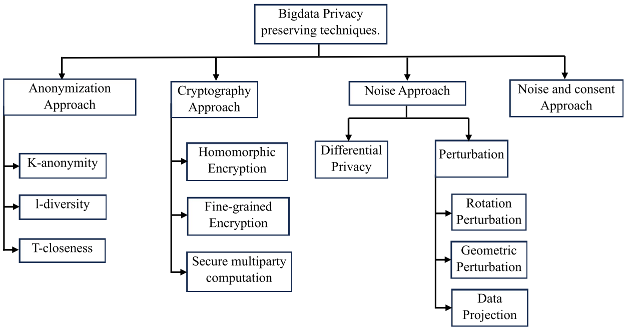

FIGURE 2. **Big data privacy techniques.**

and usage across various devices [19], [20], [21], [22]. As shown in Figure 1, the initial category addresses Infrastructure security, which revolves around thwarting potential attacks by securing both the mappers and the data in the presence of untrusted mappers. Additionally, this category delves into safeguarding NoSQL databases. Moving on, the second category, Data privacy, examines the utilization of traditional privacy-preserving techniques in data mining and analytics, such as cryptographic methods for the protection of sensitive and personal data. It emphasizes data-centric security and granular access control enforcement. The third category, Data Management, focuses on auto-tiering solutions, which pose challenges in securing data storage, emphasizing the significance of audit information for comprehending events and errors, and dealing with the intricacies of Big Data provenance. Finally, the fourth category, Integrity, and Reactive Security, deals with the validation of streaming data through endpoint validation and filtering, as well as the prevention of security issues through real-time security monitoring and analytics.

These four key domains collectively underscore the multifaceted nature of security challenges within the realm of Big Data, with each domain demanding dedicated attention and innovative solutions to ensure the integrity, privacy, and reliability of vast and diverse datasets [23].

On the other hand, privacy stands out as the foremost concern amidst various security considerations. The essence of Big Data privacy lies in its mission to shield sensitive and personal data within the vast and diverse landscape of information. Traditional privacy methods, such as cryptography, prove inadequate for the unique privacy demands of Big Data, resisting direct application. This has prompted a pressing need for effective privacy strategies tailored to the preservation of Big Data confidentiality [24], [25]. As illustrated in Figure 2, a range of techniques address Big Data privacy concerns, including those discussed in [16], [18], and [26]. Moreover, Table 1 offers a comparative analysis of these techniques. Here are some important privacy techniques employed in the context of Big Data:

A. ANONYMIZATION APPROACH

Anonymization is a critical privacy technique used to protect the confidentiality of data by removing or obfuscating identifying information. Anonymization approaches are employed to ensure that individuals cannot be re-identified from the data, even when combined with external information or advanced techniques. Here are some approaches depend on Anonymization: K-anonymity, L-diversity, T-closeness.

Firstly, K-anonymity approach ensures that each record in the dataset is indistinguishable from at least K-1 other records. This means that an individual's data cannot be singled out without ambiguity [27], [28], [29]. K-anonymity is attained through a combination of suppression and generalization techniques. In suppression, specific attribute values are substituted with a symbol *, while generalization involves replacing individual attribute values with broader categories. For instance, an attribute like *Age* with a value of 28 might be replaced with larger than 30, and a value of age could become larger than 10 and less than or equal 20.

Secondly, L-Diversity: Go beyond K-anonymity by ensuring that the sensitive attribute (e.g., a disease diagnosis) has at least L distinct values within each group of K-anonymous records. This adds an additional layer of privacy protection.

Thirdly, T-Closeness: Guarantee that the distribution of sensitive attributes within a group is not significantly different from the overall dataset's distribution. This prevents attackers from making inferences about rare or unusual values.

B. CRYPTOGRAPHY APPROACH

The cryptography approach is a fundamental technique used to enhance data security and privacy in various domains, including information technology, communication, and data storage. This approach involves the use of cryptographic methods and algorithms to protect data from unauthorized access, disclosure, or tampering. There are some approaches of cryptography such as: Homomorphic Encryption, fine-grained encryption, and Secure multiparty computation [30].

As outlined in references [5], the Homomorphic encryption technique empowers computations to be executed on encrypted data, all without the necessity of decrypting it, thereby upholding data privacy. In contrast to the Fine-Grained Encryption Method (FGEM), which selectively shields only vital data components through various encryption algorithms, thereby curbing execution time and preserving system efficiency, researchers in [31] have ingeniously harnessed a tree structure for data storage. This approach facilitates operational procedures aimed at isolating critical segments for FGEM application.

Furthermore, in the works of [27] and [32], astute protection of extensive big data repositories is accomplished by deploying pointers to safeguard pivotal sections within unstructured and semi-structured data. FGEM, in this context, encrypts the addresses of these pointers. This fine-grained methodology contributes significantly to the reduction of encryption and decryption processes, culminating in heightened system speed and efficiency. Conversely, the secure multiparty computation technique engages in secure multiparty computations, wherein each participant contributes

TABLE 1. Comparison between different privacy techniques.

Techniques	Features	Advantages	Drawback	Data utility
k-anonymity	Combining groups of data with related characteristics makes it possible to hide the identity of any one of the people who contributed to the data.	K-anonymity creates a huge equivalence class to avoid record linking.	The compromised attacker can nevertheless relate the sensitive value of a person without knowing his record if most records in an equivalence class have Utility may be similar values on a sensitive property.	Utility may be weakened such that every query produces at least k matches.
l-Diversity	Based on the idea that diversity within a group precludes the variety of sensitive values	l-Diversity protects against homogeneity attacks and knowledge attacks, among other things.	It may be challenging to establish, unnecessary, and insufficient to avoid attribute disclosure.	No usefulness for data
t-closeness	protects data privacy by reducing the level of detail in a data representation.	t-Closeness addresses the diversity-related vulnerabilities related to attribute disclosure: Attack using skewness and similarity	It limits the amount of useful information that is released. It destroys the correlations between key and confidential attributes	Utility is damaged when t is very small
Homorophic encryption	The user will be able to utilize the encrypted text and carry out various operations without having to decrypt it.	organizations may raise the bar for data security without jeopardizing application functioning or business operations.	When computing on ciphertext, there is a significant computational cost and complexity.	forces a trade-off between value and data privacy
Differential privacy	allows to compare stored information in databases with external data to obtain meaningful information from those databases without compromising the privacy of the individuals.	Most appropriate for huge data. offers the highest assurance of privacy.	The probability of attacking the database by adversary is reduced not taken into consideration.	Data utility may be reduced.
Perturbation	Changing the value of the data without altering the meaning of it	Various independently maintained characteristics	There is no way to recreate the original data values.	No data utility
Notice and consent	makes the user's private rights guaranteed, allowing him to preserve his privacy.	an internet-specific privacy technology	Applying it presents a number of difficulties.	Helps in data utility
Proposed technique	It is secure and efficient strategy to protect Big data privacy. used an encryption technique to encrypt the sensitive data and add a small amount of noise to quasi-identifier to change the form of the other data	It increases the speed and efficiency of system and maintains big data privacy	There are no drawbacks yet.	creates a tradeoff between data privacy and utility. improve the high level of data utility.

inputs geared towards a common goal [25], [26]. This methodology primarily addresses challenges inherent to function computations involving distributed inputs. It prioritizes party security, emphasizing key security attributes such as privacy and accuracy. Pertaining to privacy within the secure protocol, the principle dictates that no information beyond the function's output be disclosed, thereby meticulously preserving confidentiality.

In general, The cryptography approach is a cornerstone of modern data security and privacy. It provides the means to protect data from various threats, including unauthorized access, eavesdropping, and data breaches. However, it is essential to use strong encryption algorithms, keep cryptographic keys secure, and stay informed about emerging cryptographic techniques to maintain effective data protection.

C. NOISE-BASED APPROACH

The Noise-Based Approach is a privacy-preserving technique used in data analytics and machine learning to protect sensitive information while still extracting meaningful insights from the data. It involves adding controlled random noise to the data in order to obscure individual details while preserving the overall statistical properties of the dataset. This approach is particularly important in situations where data needs to be shared or analyzed by third parties while ensuring the privacy of individuals or entities within the data [10], [33].

In the Noise-Based Approach, it becomes challenging for adversaries to pinpoint the specific target data due to the deliberate introduction of noise. Notable methodologies employed in this approach encompass Differential Privacy and Perturbation techniques, such as rotation and geometric projection. Differential Privacy, as outlined in references [9], [10], empowers analysts to extract valuable insights from databases containing sensitive personal information without compromising individuals' privacy. This is achieved through the deployment of specialized software, often referred to as a ''privacy guard,'' which serves as a protective barrier between the database and users.

The privacy guard software allows users to submit queries to the database, evaluates these queries, assesses privacy risks, and subsequently furnishes the query results from the database to users with the incorporation of additional noise.

In the realm of Perturbation techniques, the work of [26] takes center stage. Perturbation techniques ensure data privacy by altering the values of data records while retaining their inherent meaning. This approach harnesses two distinct methods: probability distribution, which entails replacing data values either with themselves or others from the same distribution, and value distortion, achieved by introducing additive or multiplicative noise to the original data [34].

Perturbation encompasses three primary types: 1) Rotation Perturbation: This method involves the rotation of values for pairs of attributes within the data matrix without altering their underlying meaning. Value distortion techniques are subsequently applied to these attribute pairs after selection [26]. 2) Geometric Perturbation: An enhancement of rotation perturbation, geometric perturbation introduces additional components as noise alongside the basic multiplicative perturbation form represented as Y = R £X [35]. 3) Data Projection: This technique involves projecting a set of data points from a high-dimensional space to a randomly selected lower-dimensional space. This process transforms the original data into a perturbed state while preserving many of its distance-related characteristics [36].

In [25], the authors introduced the Horus technique, which employs keyed hash trees (KHTs) to encrypt extensive datasets, offering a nuanced approach to security. This method entails generating unique keys for distinct dataset regions, ensuring that each key is isolated from accessing data in other regions. Subsequently, they developed a system utilizing the firework algorithm to obfuscate sensitive information [31]. Their approach involves the creation of new pixels from selected ones (referred to as "fireworks") to replace existing pixels. While they proposed the utility of a new Firework Algorithm (FWA) for enhancing security aspects, the focus was not on optimization problems. This system encompasses three rotation-based perturbation methods for image data, aimed at concealing fine details by manipulating pixel orientations in images. The first approach utilizes the firework algorithm, placing fireworks on selected pixels and expressing the rotation processes through sparks. The second method employs a rotation-based perturbation model that integrates the iterated local search algorithm (ILS) with two optimization stages. In contrast, the third approach adheres to the same underlying principle but extends it by utilizing three optimization stages with the ILS algorithm.

To introduce a concept of combined partial image encryption, the authors leverage phase manipulation and sign encryption. This partial encryption is executed in two stages: initially, the image is scrambled by applying the Inverse Fourier Transform to yield a modified image. Subsequently, sign encryption is utilized to secure the sign bits of the modified image. The end result is a partially encrypted image, with Sign Encryption playing a crucial role in extracting the sign bits of the modified image [37].

In the work [38], the authors introduced a Partial Encryption Algorithm that employs a Random Number Generator to partially encrypt data, with a crucial requirement that the bit positions of multimedia messages be randomly determined. This technique empowers message senders to haphazardly select bit positions within multimedia messages once they have been converted into binary form. These selected bit positions are then encrypted using a symmetric key cryptosystem and transmitted alongside the remaining bits of the multimedia message to the recipient. Importantly, without knowledge of the encrypted bit order, decryption becomes infeasible. Additionally, Chunguang [39] proposed the fine-grained encryption method (FGEM), which specifically targets the encryption of critical data components using a chosen encryption algorithm. FGEM employs a tree structure to store data and employs various operational procedures to extract the critical components suitable for FGEM application. Subsequently, to accommodate the protection of sizable data, particularly semi-structured and unstructured data, FGEM was enhanced [32]. This enhancement involves using a pointer to identify critical sections of unstructured and semi-structured data, applying FGEM to encrypt the pointer's address using an encryption algorithm, and converting sensitive data from the dataset into numerical values [35]. Following this, a perturbation technique is employed to obscure the numerical results, and the perturbed numbers are then converted back into string values. Expanding upon FGEM, the authors in [33] introduced the fine-grained encryption perturbation technique (FGEPT) by amalgamating FGEM with a novel perturbation approach. Initially, FGEM is used to encrypt key identifiers and sensitive attributes. Subsequently, the proposed perturbation technique is applied to introduce a slight degree of random noise to quasi-identifiers, thereby thwarting data disclosure and preserving privacy.

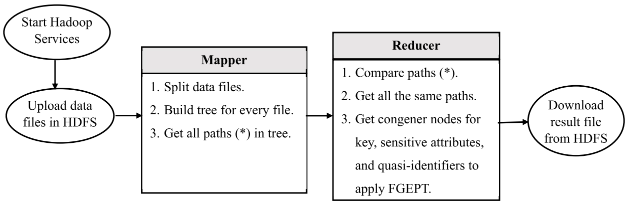
Start Hadoop Services
Upload data files in HDFS
Mapper
1. Split data files.
2. Build tree for every file.
3. Get all paths (*) in tree.
Reducer
1. Compare paths (*).
2. Get all the same paths.
3. Get congener nodes for key, sensitive attributes, and quasi-identifiers to apply FGEPT.
Download result file from HDFS

FIGURE 3. Steps of the proposed approach using MapReduce.

In approach [33], a perturbation-based encryption technique is implemented by incorporating garbage values as noise in two distinct parts of the dataset: odd and even rows. In odd-numbered rows, the noise is represented as a multiple of a designated garbage value, while in even-numbered rows, the noise takes the form of an addition of the same garbage value. Following the introduction of these garbage values into the dataset, encryption is employed using the AES encryption algorithm.

In a recent development [26], researchers presented an integrated big data system designed to support multi-objective resource optimization through fine-grained instance-level modeling and optimization. This innovative approach involves the creation of fine-grained instance-level models capable of encoding all relevant information as multi-channel inputs to deep neural networks. While the authors in [40] introduced a novel fine-grained access control system, addressing issues related to access control, user computing efficiency, and privacy protection within cloud storage services. Our system offers the capability to restrict users' access times and demonstrates provable security under a newly established security model rooted in simulation. At the heart of our system lies a new CP-ABE (Ciphertext-Policy Attribute-Based Encryption) scheme that facilitates verifiable outsourced decryption.

III. PROPOSED TECHNIQUE

Given the inherent nature of Big Data, one of its foremost challenges revolves around security and privacy concerns. The amalgamation of data from diverse sources for analysis, storage, and management inherently exposes the data to potential vulnerabilities, thus accentuating the criticality of security and privacy. Therefore, an effective privacy mechanism is imperative to uphold the confidentiality of Big Data.

In this paper, we introduce a robust and efficient strategy dedicated to safeguarding Big Data's privacy. Our approach enhances the Fine-Grained Encryption Method (FGEM) by incorporating a novel perturbation technique designed to bolster Big Data's privacy defenses. This involves the encryption of sensitive data and the introduction of minimal noise to quasi-identifiers, thereby altering the structure of other data. The Hadoop framework has been completely utilized to implement the proposed technique, which has been specifically employed in the analysis of healthcare data. This dataset encompasses various data types, such as PatientID (characters), Patientname (characters), PatientGender (characters), PatientAddress (characters), PatientAge (integer), PatientHistory (characters), and so on.

In general, Hadoop provides a distributed file system for managing large-scale datasets. It employs the MapReduce programming model to convert these datasets into valuable information. Big datasets typically span hundreds of gigabytes, necessitating a distributed file system like Hadoop Distributed File System (HDFS). This allows the data to be stored across clusters of diverse commodity machines, enabling parallel access. MapReduce operates through two distinct stages: the Map phase and the Reduce phase. During the Map phase, a dataset is segmented into key-value pairs.

The resultant output from the Map phase serves as input for the Reduce phase, where it undergoes reduction to form smaller key-value pairs. Ultimately, the key-value pairs generated by the Reduce phase constitute the final output of the MapReduce process. It's important to emphasize that the Reduce phase occurs only upon the completion of the Map phase. Until then, the Reduce phase remains inactive or blocked. The primary advantage of the MapReduce paradigm lies in its capability to enable parallel processing of data across a vast cluster of commodity machines. This parallelism significantly enhances output efficiency, delivering higher results within a shorter timeframe. As depicted in Figure 3, the key steps of our approach are outlined as follows:"

1) Initiate the Hadoop platform to implement the proposed technique on large-scale data.
2) Begin storing the extensive dataset within the Hadoop Distributed File System (HDFS) to facilitate accessibility.
3) Employ the MapReduce framework to initiate the application of the proposed technique as follows:
   a) Initially, the mapper divides the dataset file into separate files, reducing processing time for large Big Data files.
   b) Subsequently, it constructs a tree for each file, preserving patient-related values.
   c) This approach accelerates the search for paths to related nodes associated with key, sensitive, and quasi-identifiers compared to searching through the entire file.
   d) Definitions such as Label path, data path, path instance, and congener nodes are extracted from the tree structure, facilitating the isolation of specific data segments for the application of the proposed technique.
   e) Utilizing the constructed tree, the mapper identifies key attributes (e.g., id, name), sensitive attributes (e.g., address, history), and quasi-identifiers (e.g., age, gender), as illustrated in Table 2.
   f) The mapper processes these attributes as key-value pairs, generating paths for each patient and storing values within nodes for subsequent processing by the reducer.
   g) In the next step, the Reducer scrutinizes every path, focusing on key attributes, sensitive attributes, and quasi-identifiers to identify matching paths and extract congener nodes.
   h) These congener nodes are then retrieved from the paths, organized into blocks, and their values are stored within these blocks.
   i) Following this, the reducer manipulates key-sensitive attributes and quasi-identifiers from the blocks, applying the proposed technique by encrypting key and sensitive attributes and introducing noise to the quasi-identifiers.
   j) Ultimately, the result data is substituted within the file, and the initially divided input files are consolidated to produce a de-identified and more secure file, thus completing the implementation of the proposed technique.

IV. DISCUSSION AND RESULTS

The experimental setup involves the implementation of the proposed technique, employing various encryption methods, on diverse scales of healthcare Big Data across different computing platforms. The encryption techniques employed in our implementation include RSA (Rivest-Shamir-Adleman)

TABLE 2. Attribute type.

ATTRIBUTE TYPE	PROPERTY	EXAMPLE	ACTION REQUIRED
KEY	Can identify an individual directly	id, name	Fine grained encryption
QUASI IDENTIFIER	Can be linked with external information to identify an individual	age, gender	Add noise
SENSITIVE ATTRIBUTE	Data that an individual is sensitive about revealing	address, history	Fine grained encryption
UN-SENSITIVE ATTRIBUTE	include those attributes which are not sensitive.	condition, suit	No action

and Data Encryption Standard (DES). The RSA encryption algorithm, a prevalent asymmetric encryption method, finds extensive application across various products and services. It relies on a linked mathematical key pair for data encryption and decryption. It involves generating a private and a public key, with the public key accessible to anyone and the private key known only to the key's creator. This dual-option feature of encryption provides versatility to RSA users. Using the public key for encryption mandates the private key for decryption. This proves beneficial for secure data transmission across networks or the internet. In this scenario, the data sender uses the recipient's public key to encrypt sensitive information before transmitting it. As only the owner of the private key can decrypt data encrypted with the public key, the intended recipient remains the sole entity capable of deciphering it, even if intercepted during transit [41].

The DES algorithm operates as a symmetric-key block cipher. It processes plain text in 64-bit blocks and transforms it into ciphertext utilizing 48-bit keys. Unlike the RSA algorithm, DES uses a single key for both encrypting and decrypting the data [42].

Initially, we employ FGEPT with the RSA encryption algorithm to gauge execution time across distinct platforms, as illustrated in Figure 4. Notably, during the first platform execution (absent Hadoop), the process halted at a specific data size due to insufficient memory capacity. Subsequently, in the second platform execution, we apply FGEPT with the RSA encryption algorithm within a single-node Hadoop cluster environment. It's worth noting that in the single-node Hadoop cluster setup, all daemons (including DataNode, NameNode, TaskTracker, and JobTracker) operate on the same machine or host. However, due to memory limitations, execution came to a halt at a specific data size, surpassing the data size threshold observed in the first platform without Hadoop. Furthermore, we observed that the utilization of FGEPT with the RSA encryption algorithm within the single-node Hadoop cluster incurred a lower execution time compared to the initial platform without Hadoop. Moving forward, we extended our experimentation to a multi-node Hadoop cluster platform, characterized by a main-secondary architecture.

In this configuration, the main node hosts the NameNode daemon, while multiple secondary nodes, distributed across different machines or hosts, run DataNode, TaskTracker, and JobTracker daemons. Interestingly, as the dataset size increased, we observed that the execution time in the multi-node Hadoop platform consistently outperformed that in the other platforms. As indicated in Figure 4 and corroborated by Figure 9, FGEPT using the RSA encryption algorithm within the multi-node Hadoop platform exhibited a notable reduction in execution time, amounting to approximately 14% less compared to other platforms.

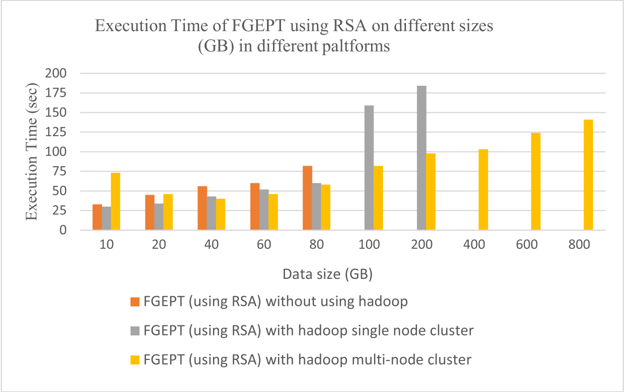

FIGURE 4. Execution time of FGEPT using RSA in different platforms.

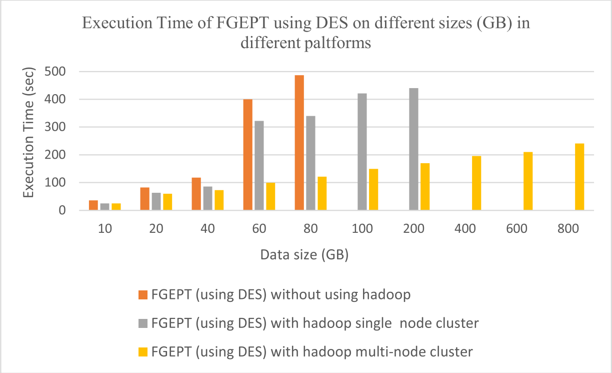

FIGURE 5. Execution time of FGEPT using DES in different platforms.

Next, as depicted in Figure 5, we proceed to implement FGEPT, this time employing the DES encryption algorithm, to evaluate execution times across various platforms. Notably, during the initial platform execution (absent Hadoop), the execution encountered memory limitations, causing it to halt at a specific data size.

In the subsequent platform execution, we leverage FGEPT with the DES encryption algorithm within a single-node Hadoop cluster. It's worth highlighting that this execution also experienced a termination point at a data size threshold exceeding that of the first platform. Moreover, we observed that FGEPT utilizing the DES encryption algorithm within

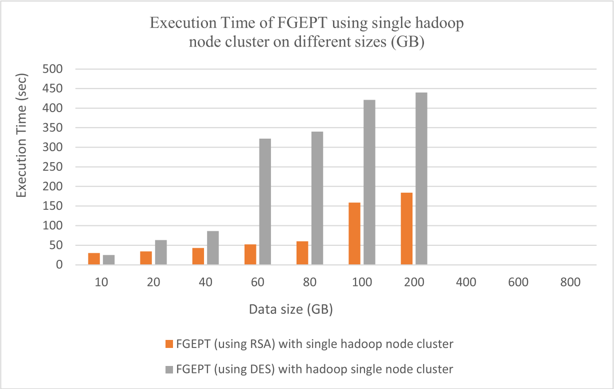

FIGURE 6. Execution time of FGEPT using single Hadoop node cluster.

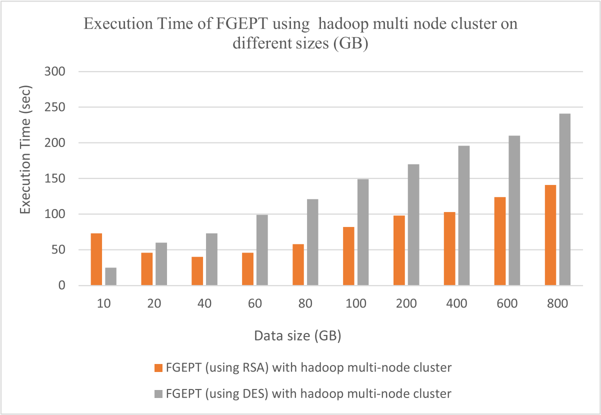

FIGURE 7. Execution time of FGEPT using Hadoop multi-nodes cluster.

the single-node Hadoop cluster exhibited a significant reduction in execution time, approximately 48% less compared to the initial platform without Hadoop. The reason for this is that utilizing FGEPT with a single-node Hadoop system involves task division for parallel execution. Consequently, this parallel processing simplifies task allocation, ultimately resulting in quicker program execution.

Lastly, we extended our investigation to the multi-node Hadoop cluster platform, and Figure 5 reveals that FGEPT with the DES encryption algorithm in this configuration delivered superior results compared to FGEPT with the RSA encryption algorithm in other platforms. Additionally, Figure 6 offers a comparison between FGEPT utilizing RSA and FGEPT utilizing DES within a single-node Hadoop cluster platform.

As illustrated in Figure 8, we observe that FGEPT utilizing the RSA encryption algorithm significantly reduces execution time, with an 8% decrease compared to FGEPT employing the DES encryption algorithm, which incurs a 48% reduction. Simultaneously, FGEPT with the DES encryption algorithm exhibits a higher capacity for processing larger data sizes than FGEPT with the RSA encryption algorithm. Furthermore, we conducted a comparative analysis between FGEPT utilizing RSA and FGEPT using DES within the multi-node Hadoop cluster platform, as shown in Figure 7. According to Figure 8, it's evident that FGEPT employing the RSA encryption algorithm reduces execution time by approximately 14% when contrasted with FGEPT using the DES encryption algorithm, which experiences a 30% reduction.

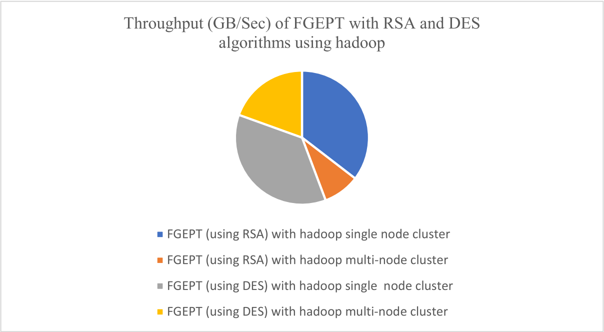

FIGURE 8. Throughput of FGEPT with RSA and DES using Hadoop multi-nodes cluster.

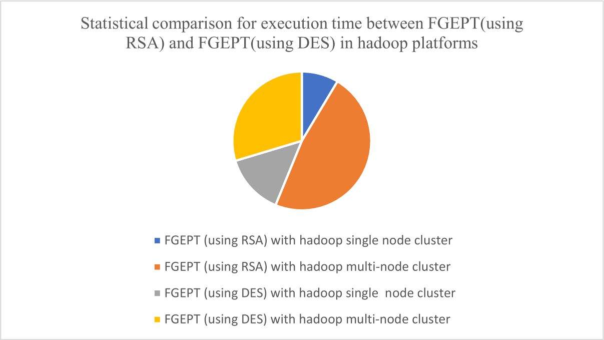

FIGURE 9. Statistical comparison for execution time of FGEPT using RSA and DES algorithms in Hadoop platform.

In addition, Figure 6 and Figure 7 highlight that FGEPT with the DES encryption algorithm possesses the capability to process larger data sizes than FGEPT utilizing the RSA encryption algorithm.

To conclude, we determine the throughput of the encryption algorithm in both the Hadoop single-node cluster and Hadoop multi-node cluster platforms by calculating the average plain text size in megabytes divided by the average execution time in seconds.

As depicted in Figure 9, it's worth noting that FGEPT, when utilizing the RSA encryption algorithm within both the Hadoop single-node cluster and Hadoop multi-node cluster platforms, exhibits a higher throughput compared to when employing the DES encryption algorithm. Consequently, FGEPT employing the DES encryption algorithm consumes less power in contrast to its RSA counterpart. Turning to Figure 8, we observe that within the Hadoop execution platform, FGEPT employing the RSA encryption algorithm attains the most favorable results in terms of execution time. Finally, when assessing the speedup achieved by the proposed technique for RSA and DES encryption, we find the following results: The speedup for the proposed technique employing the RSA encryption technique is equal to 1.04 Ms, while the speedup for it utilizing the DES encryption technique equals 2.66 MS.

V. CONCLUSION

Security represents the paramount challenge within the realm of Big Data. Traditional security approaches, particularly those pertaining to privacy, such as cryptography and anonymization, have demonstrated inadequacies in addressing the unique demands posed by Big Data. In this study, our primary objective revolves around fortifying the privacy of Big Data by enhancing the Fine-Grained Encryption Method (FGEM) to align with the specific requisites of Big Data environments. Our proposed technique is rigorously assessed across various scales of Big Data. The experimental findings unequivocally highlight the superior performance of our approach when employing the RSA encryption algorithm within the Hadoop execution platform, particularly in terms of memory usage and execution time. In summary, our approach not only augments privacy safeguards but also preserves a high level of data utility by effectively obviating the potential linkage between Big Data and external datasets.

ABBREVIATIONS AND ACRONYMS

The following abbreviations are used in this manuscript:

FGEM	Fine Grained Encryption Method
KHTs	key hash trees
FWA	firework algorithm
FWA-IP	firework algorithm-image perturbation
ILS-IP	local search algorithm-image perturbation
AES	Advanced Encryption Standard
RSA	Rivest-Shamir-Adleman
HDFS	Hadoop Distributed File System

ACKNOWLEDGMENT

The authors would like to acknowledge the support of Imam Abdulrahman Bin Faisal University, especially the Department of Computer Science, Applied College.

REFERENCES

[1] S. P. Arockia, S. S. Varnekha, and A. K. Veneshia, "The 17 vs of big data," *Int. Res. J. Eng. Technol. (IRJET)*, vol. 4, no. 9, pp. 329–333, 2017.

[2] A. A. Hussien, "Fifty-six big data vs characteristics and proposed strategies to overcome security and privacy challenges (BD2)," *Int. J. Inf. Secur.*, vol. 11, no. 4, pp. 304–328, 2020.

[3] A. Rahmani, A. Amine, and R. H. Mohamed, "A multilayer evolutionary homomorphic encryption approach for privacy preserving over big data," in *Proc. Int. Conf. Cyber-Enabled Distrib. Comput. Knowl. Discovery*, Shanghai, China, Oct. 2014, pp. 19–26.

[4] C. Butpheng, K. H. Yeh, and H. Xiong, "Security and privacy in IoT-cloud-based e-health systems—A comprehensive review," *Int. J. Symmetry*, vol. 12, no. 7, pp. 1–35, 2020, doi: 10.3390/sym12071191.

[5] R. Bao, Z. Chen, and M. S. Obaidat, "Challenges and techniques in big data security and privacy: A review," *Secur. Privacy*, vol. 1, no. 4, p. e13, Jul. 2018, doi: 10.1002/spy2.13.

[6] N. Odugu and A. Rajesh, "A fine-grained access control survey for the secure big data access," *Int. Turkish J. Comput. Math. Educ.*, vol. 12, no. 10, pp. 4180–4186, 2021.

[7] E. A. Elwan, M. Elkawkagy, and A. Keshk, "Enhancing fine grained technique for maintaining data privacy," *Int. J. Adv. Res. Comput. Appl.*, vol. 10, no. 1, pp. 1–6, 2018.

[8] L. El Haourani, A. A. El Kalam, and A. A. Ouahman, "Big data security and privacy techniques," in *Proc. 3rd Int. Conf. Netw., Inf. Syst. Secur.*, Mar. 2020, pp. 1–9.

[9] X. Hu, M. Yuan, J. Yao, Y. Deng, L. Chen, Q. Yang, H. Guan, and J. Zeng, "Differential privacy in Telco big data platform," *Proc. VLDB Endowment*, vol. 8, no. 12, pp. 1692–1703, Aug. 2015.

[10] P. Jain, M. Gyanchandani, and N. Khare, "Big data privacy: A technological perspective and review," *J. Big Data*, vol. 3, no. 1, pp. 1–25, Dec. 2016.

[11] T. Revathi and N. Ramaraj, "Data privacy preservation using data perturbation techniques," *Int. J. Soft Comput. Artif. Intell.*, vol. 5, no. 2, pp. 10–12, 2017.

[12] J. Koo, G. Kang, and Y. G. Kim, "Security and privacy in big data life cycle: A survey and open challenges," *Int. J. Sustainability*, vol. 12, no. 24, p. 10571, 2020, doi: 10.3390/su122410571.

[13] Y.-H. Chun and M.-K. Cho, "An empirical study of intelligent security analysis methods utilizing big data," *Webology*, vol. 19, no. 1, pp. 4672–4681, Jan. 2022.

[14] C. Zhou, C. Ma, and S. Yang, "An improved fine-grained encryption method for unstructured big data," in *Proc. Int. Conf. Young Comput. Scientists, Eng., Educators*, Harbin, China, Jan. 2015, pp. 361–369.

[15] I. P. Het, M. P. Shivani, and J. S. Ketan, "Security and privacy with perturbation based encryption technique in big data," *Int. J. Comput. Sci. Commun.*, vol. 7, no. 1, pp. 205–209, 2016.

[16] A. Rahmani, A. Amine, R. M. Hamou, M. E. Rahmani, and H. A. Bouarara, "Privacy preserving through fireworks algorithm based model for image perturbation in big data," *Int. J. Swarm Intell. Res.*, vol. 6, no. 3, p. 4158, 2015.

[17] D. Florea and S. Florea, "Big data and the ethical implications of data privacy in higher education research," *Int. J. Sustainability*, vol. 12, no. 20, pp. 1–11, 2020, doi: 10.3390/su12208744.

[18] K. N. Vachhani and D. Vaghela, "A servey on geometric data transformation for privacy preserving on data stream," *Int. J. Tech. Res. Appl.*, vol. 3, no. 2, pp. 257–259, 2015.

[19] M. Tota and C. A. Dhote, "A new approach to big data challenges and opportunities," in *Proc. Int. Conf. Innovtive Trends Eng., Sci., Manag.*, Dubai, United Arab Emirates, Dec. 2016, pp. 150–157.

[20] S. Sahu and Y. Dhote, "A study on big data: Issues, challenges and applications," *Int. J. Innov. Res. Comput. Commun. Eng.*, vol. 4, no. 6, pp. 10611–10616, 2016.

[21] A. Z. Faroukhi, I. El Alaoui, Y. Gahi, and A. Amine, "A multi-layer big data value chain approach for security issues," *Proc. Comput. Sci.*, vol. 175, pp. 737–744, Jan. 2020.

[22] L. Sun, H. Zhang, and C. Fang, "Data security governance in the era of big data: Status, challenges, and prospects," *Data Sci. Manag.*, vol. 2, pp. 41–44, Jun. 2021.

[23] S. Salini, S. V. Kumar, and R. Neevan, "Survey on data privacy in big data with K-anonymity," *Int. J. Innov. Res. Comput. Commun. Eng.*, vol. 3, no. 5, pp. 3765–3771, Jun. 2015.

[24] K. M. P. Shrivastva, M. A. Rizvi, and S. Singh, "Big data privacy based on differential privacy a hope for big data," in *Proc. Int. Conf. Comput. Intell. Commun. Netw.*, Bhopal, India, Nov. 2014, pp. 776–781.

[25] F. H. Cate and V. Mayer-Schonberger, "Notice and consent in a world of big data," *Int. Data Privacy Law*, vol. 3, no. 2, pp. 67–73, May 2013.

[26] C. Lyu, Q. Fan, F. Song, A. Sinha, Y. Diao, W. Chen, L. Ma, Y. Feng, Y. Li, K. Zeng, and J. Zhou, "Fine-grained modeling and optimization for intelligent resource management in big data processing," in *Proc. Int. Conf. Very Large Data Base Endowment (VLDBE)*, 2022, vol. 15, no. 11, pp. 1–34.

[27] S. G. Rizzo, F. Bertini, and D. Montesi, "Fine-grain watermarking for intellectual property protection," *EURASIP J. Inf. Secur.*, vol. 2019, no. 1, pp. 1–20, Dec. 2019.

[28] C. Eyupoglu, M. Aydin, A. Zaim, and A. Sertbas, "An efficient big data anonymization algorithm based on chaos and perturbation techniques," *Entropy*, vol. 20, no. 5, p. 373, May 2018.

[29] J. Guo, M. Yang, and B. Wan, "A practical privacy-preserving publishing mechanism based on personalized K-anonymity and temporal differential privacy for wearable IoT applications," *Int. J. Symmetry*, vol. 13, no. 6, p. 1043, 2021, doi: 10.3390/sym13061043.

[30] J. Moura and C. Serro, "Security and privacy issues of big data," *Int. J. Netw. Secur. Appl.*, vol. 8, no. 1, pp. 59–79, 2019.
[31] S. Qi, Y. Lu, W. Wei, and X. Chen, "Efficient data access control with fine-grained data protection in cloud-assisted IIoT," *IEEE Internet Things J.*, vol. 8, no. 4, pp. 2886–2899, Feb. 2021.
[32] D. J. Patel and S. Pate, "A survey on data perturbation techniques for privacy preserving in data mining," *Int. J. Sci. Res. Develop.*, vol. 3, no. 1, pp. 52–54, 2015.
[33] Y. Li, N. S. Dhotre, Y. Ohara, T. M. Kroeger, E. Miller, and D. D. E. Long, "Horus: Fine-grained encryption-based security for large-scale storage," in *Proc. 11th Int. Conf. File Storage Syst.*, Santa Clara, CA, USA, Feb. 2013, pp. 147–160.
[34] L. Patel and R. Gupta, "A survey of perturbation technique for privacy-preserving of data," *Int. J. Emerg. Technol. Adv. Eng.*, vol. 3, no. 6, pp. 52–59, 2015.
[35] P. J. Krishna and M. A. Geetha, "Perturbation of string values," *Int. J. Comput. Sci. Inf. Technol.*, vol. 2, no. 3, pp. 1257–1259, 2011.
[36] Z. Ai, Y. Liu, L. Chang, F. Lin, and F. Song, "A smart collaborative authentication framework for multi-dimensional fine-grained control," *IEEE Access*, vol. 8, pp. 8101–8113, 2020.
[37] B. D. Parameshachari, K. M. S. Soyjaudah, and K. A. Sumitra, "Secure transmission of an image using partial encryption based algorithm," *Int. J. Comput. Appl.*, vol. 63, no. 16, pp. 33–36, Feb. 2013.
[38] P. Agrawal and M. Rajpoot, "Partial encryption algorithm for secure transmission of multimedia messages," *Int. J. Comput. Sci. Technol.*, vol. 3, no. 1, pp. 467–470, 2012.
[39] C. Ma, C. Zhou, W. Jiuru, and X. Zhong, "A research of fine-grained encryption method for IoT," *J. Comput. Inf. Syst.*, vol. 8, no. 24, pp. 10213–10222, 2012.
[40] X. Liu, H. Wang, B. Zhang, and B. Zhang, "An efficient fine-grained data access control system with a bounded service number," *Inf. Sci.*, vol. 584, pp. 536–563, Jan. 2022.
[41] R. Imam, Q. M. Areeb, A. Alturki, and F. Anwer, "Systematic and critical review of RSA based public key cryptographic schemes: Past and present status," *IEEE Access*, vol. 9, pp. 155949–155976, 2021, doi: 10.1109/ACCESS.2021.3129224.
[42] R. Ramya Devi and V. Vijaya Chamundeeswari, "Triple DES: Privacy preserving in big data healthcare," *Int. J. Parallel Program.*, vol. 48, no. 3, pp. 515–533, Jun. 2020, doi: 10.1007/s10766-018-0592-8.

MOHAMED ELKAWKAGY received the Ph.D. degree from the Faculty of Engineering, Ulm University, Germany, in 2011. From 2011 to 2013, he was a Postdoctoral Researcher with the Artificial Intelligence Research Group, Ulm University. Since November 2016, he has been a Staff Member of the Department of Computer Sciences, Applied College, Imam Abdulrahman Bin Faisal University, Saudi Arabia, where he is currently an Associate Professor in artificial intelligence. His research interests include artificial intelligence, cloud computing, security, and big data.

E. ELWAN, photograph and biography not available at the time of publication.

ALBANDARI ALSUMAYT received the master's degree in computer security from The University of Manchester, in 2012, and the Ph.D. degree in computer security and networks from Nottingham Trent University, in 2017. Her research interests include security, wireless networks security, drones, and blockchain.

HEBA ELBEH received the Ph.D. degree from the Faculty of Engineering, Ulm University, Germany, in 2012. Since November 2016, she has been a Staff Member with the Department of Computer Sciences, Applied College, Imam Abdulrahman Bin Faisal University, Saudi Arabia, where she is currently an Associate Professor in artificial intelligence. Her research interests include artificial intelligence, blockchain, security, and big data.

SUMAYH S. ALJAMEEL received the B.S. degree (Hons.) in computer science from King Faisal University, Saudi Arabia, in 2004, the M.S. degree (Hons.) in software engineering from The University of Manchester, U.K., in 2013, and the Ph.D. degree in artificial intelligence from Manchester Metropolitan University, U.K., in 2018. She is currently an Assistant Professor of computer science and the Chair of the Department of Computer Science, College of Computer Science and Information Technology, Imam Abdulrahman Bin Faisal University, Saudi Arabia. Her research interests include machine learning, deep learning, data mining, and specifically the application of AI with other fields, such as health, energy, and oil pipeline.