Contents lists available at Science-Gate

International Journal of Advanced and Applied Sciences

Journal homepage: http://www.science-gate.com/IJAAS.html

Big data management: Security and privacy concerns

Ibrahim A. Atoum *, Ismail M. Keshta

Computer and Information Sciences Department, College of Applied Sciences, AlMaarefa University, Riyadh, Saudi Arabia

ARTICLE INFO

Article history:
Received 6 November 2020
Received in revised form
27 January 2021
Accepted 28 January 2021

Keywords:
Big data
Internet of things
Big data analytics
Big data privacy

ABSTRACT

Big data has been used by different companies to deliver simple products and provide enhanced customer insights through predictive technology such as artificial intelligence. Big data is a field that mainly deals with the extraction and systemic analysis of large data sets to help businesses discover trends. Today, many companies use Big Data to facilitate growth in different functional areas as well as expand their ability to handle large customer databases. Big data has grown the demand for information management experts such that many software companies are increasingly investing in firms that specialize in data management and analytics. Nevertheless, the issue of data protection or privacy is a threat to big data management. This article presents some of the major concerns surrounding the application and use of Big Data about challenges of security and privacy of data stored on technological devices. The paper also discusses some of the current studies being undertaken aimed at addressing security and privacy issues in Big Data.

1. Introduction

Big data defines the different sources of information that produce data such as closed-circuit television (CCTV) sensors, social media accounts, email messages, Internet of Things (IoT) devices, and different platforms that collect customer information (Wu et al., 2020). Suoniemi et al. (2020) asserted that developed countries are increasingly implementing Big Data technologies to analyze large customer databases to come up with useful insights. Big Data offer many advantages to different organizations in line with improving the decision-making processes for businesses or healthcare institutions, and improving consumer experience via customer data analysis (Atoum and Al-Jarallah, 2019). The main goal of Artificial Intelligence (AI) technology in the field of Big Data is to help in identifying and generating knowledge that can be applied to solve problems (Reis et al., 2020). According to Sardi et al. (2020), the opportunities that Big Data has to offer are endless, a few of them include the integration of technology tools that boost the efficiency of conducting business processes.

The uses of artificial intelligence and data mining as tools that promote Big Data analytics speed up the collection of customer insights aimed at improving their loyalty to the brand (Davenport et al., 2020). However, as Martin et al. (2017) viewed it, Big Data has challenges with privacy and security that affect data management by organizations as well as consumer willingness to provide data in exchange for free services.

In this paper, we will discuss the issues related to big data privacy and security, which include: Background to Big data, Big Data Trends, Security Challenges and Privacy Issues of Big Data, Big Data Infrastructure Security, Integrity and Reactive Security of Big Data, Lack of Anonymity and Data Privacy Technology, Data Management, Stakeholder Role in Big Data Security, and Proposed Security Methods.

2. Motivation and main contribution

The paper makes a summary of most of the recent approaches applied to Big data to deliver reliable privacy measures for protecting user data. The paper states methods used to mitigate Big Data risks such as the effectiveness of the MapReduce algorithm and how it has made the anonymization of big data streams quite easier. The paper also proposes simple proactive measures such as active monitoring of Big Data systems and user authentication to prevent unwarranted access to sets of data. Thirdly, the paper illustrates the efficient and effective

* Corresponding Author.
Email Address: iotoum@mcst.edu.sa (I. A. Atoum)
https://doi.org/10.21833/ijaas.2021.05.009
Corresponding author's ORCID profile:
https://orcid.org/0000-0002-9259-7937
2313-626X/© 2021 The Authors. Published by IASE.

approaches to ensuring Big Data Security and Privacy and prevention of information loss during the anonymization process. Security aspects in big data are associated with the internet bubble propelling the digitization and storage of records on online platforms.

As a result, the Big Data industry is witnessing rapid growth in data storage which introduces complexity in analysis. Analyzing big data requires tools that can agglomerate large and complex data sets. Being able to computationally guarantee data sharing and privacy in the storage and communication capabilities of Big Data requires conventional methods of system administration and security. In Big Data, various factors deliver the necessary objects to harness the power of big data. The power of big data analysis and research on client behavior through the use of computational algorithms ensures real-time access while ensuring security for user data. Likewise, companies are able to tailor their needs to be client-oriented based on rational Big Data analysis.

The main contributions of this research paper include highlighting some of the mechanisms used to address the issue of big data privacy. The paper mentions five contributions. The adoption of big data in many companies including social media and healthcare has significantly increased security and privacy concerns for the end-user.

3. Background on big data

Information privacy is an issue that has been a major challenge that threatens the advancement of Big Data. Big data simply refers to data sets that are often too large and complex for normal technological applications to process normally. The large-scale collection of data through the internet allows companies such as Facebook, Twitter, and Google platforms to gather a lot of customer information frequently. Arguably, the users who access services offered by these big companies rarely go through privacy policies. As a result, the customers are often subjected to risk when their data end up in the possession of third-party entities. The slow communication of policy preferences to technology devices and social network users remain to be a major obstacle to solve the data privacy challenge. The era of Big Data is an opportunity that both governments and businesses need to collect vital data and information to improve their customer services (Alharthi et al., 2017). Nevertheless, the clients are rarely informed that the data collected from them could be sold to third-party requests or other business entities. The challenge of Big Data management resulting in privacy and security violations is a significant concern since data can always be compromised. Hence, as organizations actively adopt the use of Big Data tools within their business processes, they need to upgrade the strategies for ensuring the privacy and security of their Big Data platforms.

4. Big data trends

Government agencies and businesses are continuously collecting and generating large data outputs. The current increased focus on big data has made it possible for companies to process and analyze the data in different business domains. Subsequently, there is a growing need for developers to create security resilient software applications to show just how Big Data has quickly evolved to center-stage. The current trends in Big Data show that its implementation in collection and performing analytics of vast amounts of data has been slowly developing over the last few decades. In the course of the past decade, big data has been improved to solve vital issues such as make conclusive findings from healthcare records that accumulated from many information transactions and storage of sensitive patient data (Viceconti et al., 2015; Atoum and Al-Jarallah, 2019). Big data seems to be evolving simultaneously with progression in technology where new concepts are being innovated while others integrate a mix of different computer technologies to improve Big Data analytics. Therefore, as we advance in the IoTs, future trends in big data will continue to grow and be a powerful tool to assist many companies. Big data is opening up the possibilities for industries to provide insights into customer-centered questions and tasks that have never been known by data scientists.

5. Security challenges and privacy issues of big data

The issue of data privacy is a risk that clients recognize when using and accessing online platforms that provide them with different services. We live in an era where Big Companies that collect customer information can predict consumer trends and make recommendations based on online behavior. However, the use of Big Data information to solve different business problems is being threatened by data privacy and confidentiality concerns. The sensitive data or information that are shared with third-party data repositories and media companies show that there is still a need to enforce data security to ensure privacy while online.

Data extraction from multiple sources is referred to as data mining, a process used to collect user data for analysis. Data mining can also be defined as the process through which organizations turn the collected raw data into knowledge through identifying patterns. In the view of Lekshmy and Rahiman (2020), there is a growing problem with data mining and big data as mismanagement during transit or storage of data that is threatening social ethics. There is no guarantee that all organizations secure the data they have on their databases. Therefore, when data brokers store customer data, they plunge their clients' privacy to risk. Some of the privacy issues related to big data include the exposure of customer details which in turn cause harm to the individuals' reputations. Privacy breaches have become sources of embarrassment to users and companies especially in line with disclosure of their social security numbers, intimate healthcare, and or private financial records (Viceconti et al., 2015). Since the technological field of big data is new to most organizations, most companies have not yet dedicated resources towards identifying and tackling risks associated with data mining. Regulators and lawmakers around the world are just beginning to acknowledge the significance of Big Data security. Reports have specified the drawbacks of Big Data Analytics; for example, unintentional discriminations based on user identities were identified as one of the reasons why laws governing Big Data Analytics ought to be made a priority.

Big Data infrastructure is increasingly becoming a challenge for most organizations because most firms do not have the security tools needed to cover users on operations outside the company's network. Big Data advancement needs adequate policies and technology compliance to fulfill the issue of data security and privacy. Given that Big Data is important to big organizations; it is only pertinent for one to question the aspects of Big Data that can impact client data in a negative way. For example, one can ask how these organizations use this data to improve the customer experience while ensuring the mobility requirements and the policies of BYOD (Bring Your Device), support the security of data (Singh, 2020). Therefore, as Big Data expands with the help of public cloud networks, traditional security solutions become less effective. Private clouds are often confined to well-defined security perimeters and networks such as firewalls and demilitarized zones (DMZs) which ensure security within the organization (Moura and Serrão, 2015). Thus, when using Big Data, security functions are necessary for the heterogeneous composition of the different hardware infrastructure, operating systems, and network domains.

Moura and Serrão (2015) opined that in complex data networks and public cloud computing environments, the abstraction capability of Software-Defined Networking (SDN) is a vital feature or component that can enable the resourceful and secure deployment of Big Data services. The heterogeneous infrastructure of public networks has the potential to invade big data privacy and security, posing a challenge to the integrity of customer data. Some of these challenges are briefly discussed below:

- Insecure Web Interface: Insecure web applications and website pages enable hackers to access data stored on client-side nodes. This can allow an attacker to exploit an administration web portal, for example, through cross-site request forgery, cross-site scripting, or by simply performing an SQL injection. Subsequently, after gaining access to the insecure web and administration interfaces, an attacker can be able to control other devices in the same network and manipulate data stored on them.
- Insufficient Authentication/Authorization: The lack of secure authentication services implemented on the web interfaces and databases can allow a computer hacker to attack weak passwords as well as to exploit the access to privileged nodes on the internet and network devices.
- Insecure Network Services: Insecure network services such as unpatched versions of network operating systems can allow an attacker to exploit weak services running on a network node. Similarly, the attacker could utilize such weak services as a pivot point to attack other computers or servers on the same network.
- Lack of Transport Encryption: The lack of proper cryptography services and encryption of data during conveyance can allow an attacker to eavesdrop on data in transit between computing devices and IT support systems.
- Privacy Concerns: The main challenge of big data security is the issue of privacy stemming from the fact that companies tend to collect private data from devices and other users' support systems and continuously fail to protect the users' data.
- Insecure Cloud Interface: Cloud computing enables and simplifies the communication between Local Area Networks (LAN) and Wide Area Networks (WAN), however, without appropriate security controls a malicious hacker can attempt multiple attacks such as password cracking or account enumeration of web portals to access sensitive data through the cloud network.
- Insecure Mobile Interface: The lack of appropriate security controls can enable a malicious hacker to make multiple attempts to illegally penetrate mobile web applications to access data or controls via the mobile interface. Some of the identified exploits include lack of data transport encryption, insufficient authentication, and account enumeration.
- Insufficient Security Configurability: The lack of proper configuration mechanisms and insecure web deployments can allow an attacker to access data or controls on an end-user device.
- Insecure Software/Firmware: Compromised software updates introduce scenarios whereby an attacker can take advantage of unauthenticated unencrypted connections to hijack network devices. This can be done by pushing malicious updates to the host computer to enable one to get sensitive information or data.
- Poor Physical Security: Lack of proper security for hardware devices can allow the attacker to gain physical access through peripheral hacking such as using the open USB ports, SD cards, or hard disks drive to unlawfully steal data in storage or directly manipulate the device's operating system.

More detailed explanations of some of the Big Data challenges are highlighted below:

6. Big data infrastructure security

Hardware security is an imperative requirement for big data users since it is the key aspect that protects sensitive data through the implementation of personalized or customizable controls. This helps in managing client data by ensuring privacy via implementing the needed strategies to monitor company devices placed within distributed networks and out of reach by the administration. Another infrastructure issue that might lead to privacy breaches includes the lack of better practices in Big Data implementation.

6.1. Secure configurability in distributed frameworks

Distributed computing is a model in which components of a software system are shared among multiple computers to improve efficiency and performance. Distributed programming frameworks harness parallelism in computation and storage to process massive amounts of data. For example, a programming model called MapReduce implements a distributed algorithm on data clusters to help the processing and generation of large data sets (Choksi et al., 2015). It usually divides the input data sets into independent segments, which are processed by map tasks in a complete and parallel manner. The concept utilized in the initial step of mapping data is the MapReduce algorithm that reads each lump of the data set then calculates and gives an output of the list of key and value pairs. In the later steps, a Reducer is used to combine the values pairs that were matched to each unique key identified then subsequently outputs the result.

In such cases, two main attack prevention measures can be implemented to secure the mapping algorithm. That entails securing the mappers and ensuring the data that might end up in the presence of an untrusted mapper is also encoded. Untrusted mappers tend to output or generate incorrect cumulative results. In financial and scientific computations, using large-scale data sets, it is unlikely to detect the extent of significant damage. For instance, in using big data for customer segmenting or targeted advertising, it is common for marketing agencies to be tasked with the analysis of the customer-related data.

The tasks that are typically involved in a MapReduce job usually split both the input and output of the job and store it in a file-system. MapReduce is considered to be faster since it is suited to process high amounts of parallel calculations over large data sets (Choksi et al., 2015). Hadoop MapReduce is a framework for processing big data in parallel, Hadoop is an open-source software framework that stores and runs data applications on large clusters of commodity hardware. The ability to provide a lot of room for storage, real-time data analysis as well as processing big data makes Hadoop an appropriate framework for the storage of diverse databases. However, the data mappers can be left intentionally or unintentionally insecure which may lead to data leakages. An example is when a mapper reveals a distinct value via the analysis of private data, hence, undermining users' privacy.

6.2. Security practices for big data storage

Another issue with the security of infrastructure in current Big Data analysis is the lack of organization between database systems. A mismatch in databases such as SQL querying databases, the software applications that perform analysis on many forms of non-SQL processing, AI, and data mining challenge the security of big data (Choksi et al., 2015). On-relational data stores (NoSQL databases) have simplified the way big companies store their diverse information. Nevertheless, this one major security challenge of big data structures. For example, the protections of data stores against NoSQL injection attacks are still not mature. Most NoSQL DBs were designed to handle several challenges presented by data segmenting and analysis, hence, security was never an integral part of NoSQL data modeling (Martinez-Mosquera et al., 2019). In most cases, developers are advised to incorporate security in the development and design phases of application interfaces as well as when building NoSQL databases.

Limited operational support is offered by NoSQL database providers for implementing it in the company's database design model. Firms that deal with big unstructured data sets stand to exponentially benefit from migrating from a traditional relational database to a NoSQL database. However, the privacy of data and its security in NoSQL databases mainly relies on external enforcing mechanisms like software engineers and database administrators associated with the database design. Such security practices make it challenging to house and process huge data volumes. Therefore, developers normally integrate security in the program middleware to lower the number of security incidents. Martinez-Mosquera et al. (2019) explained that companies also have to go through security checks set in global policies for the middleware to strengthen the NoSQL database to match Relational Databases (RDBs) without compromising the operational features.

7. Integrity and reactive security of big data

All users' data integrity is imperative to ensure the consistency of user data stored across various platforms. The various methods already recommended include end-point computer validation and real-time monitoring of the security policies to ensure the company's infrastructure complies with the set policies. A detailed description of how reactive security can be applied to big data is highlighted as follows.

7.1. End-point user validation

Many big data companies collect their data from a variety of sources, such as end-point devices. The security risk posed on the data collection process is the validation of multiple users using a single platform (Anshari et al., 2019). First, the millions of users keying in data from millions of hardware devices within the enterprise network have to be secured to prevent malicious queries by exploiting input validation attacks. Validation of inputs of data needs to be secured to protect them from malicious queries or filtering of the input data collection by an intruder. Secondly, the filtering of big data is a key concern because it becomes implausible to successfully validate and filter data continuously.

7.2. Real-time security monitoring

The physical security configuration of data networks is significant for protecting the hardware devices that store sensitive client information. Additionally, the end-user element in the monitoring of data is susceptible to the fact that humans tend to ignore security red alerts especially when they are many false positives.

Given the rapid rate and volume of data being collected, the problem of human oblivion with cross-checking security configurations on their devices might even increase with big data.

On the other hand, big data technologies can provide insights into technology applications that allow for fast processing and analysis of different types of data. This might be achieved via the implementation tools that monitor and detect anomalies in scalable security networks and providing results through analytical tools in near real-time. Most firms can gain from this opportunity as it provides real-time security analytics, although the scenarios of application may differ. For example, in the healthcare industry, big data is largely utilized to analyze different types of data (Viceconti et al., 2015). The big data technologies potentially save the tax-payer billions of shillings given that the advantage of big data analytics makes data analysis much more accurate with claim payments which reduces bad financial schemes such as insurance fraud. In this regard, the main challenge to data privacy is the authorized users that are granted access to control and manipulate the data or an IT device storing vital client data.

8. Lack of anonymity and data privacy technology

8.1. Data mining and analytics security tools

The user interfaces used by mobile phones and computer web interfaces might be compromised if much security is not put in place to prevent attacks. The customers are often used to browsing the internet oblivious of data mining and analytics from third parties. It is common for companies to collect data from various web stores and databases that their clients utilize without their knowledge or consent. Therefore, big data enables the same companies that collect the data to invade privacy through invasive marketing thereby decreasing the civil freedoms of the customers (Perera et al., 2015). The use of databanks that anonymize data is not enough to maintain user privacy and protect their sensitive data. For instance, an on-demand video streaming service in America considered being a multinational entertainment company called Netflix recently faced a problem when their users ' anonymized data were revealed through a simple correlation between the clients' IMDB scores and Netflix movie scores.

Therefore, it is significant for government policy providers and companies to establish a set of guidelines and recommendations for preventing unintentional privacy disclosures. User data collected by business and government agencies give these entities a more sense of corporate control. The drawbacks of big data are the possibility of malicious insiders taking advantage of the data without authorization or a third-party partner acquires the data and shares the datasets with other businesses interested in mining customer private information. Similarly, intelligence agencies have been at the forefront of big data collection with the help of robust algorithms that ensure the privacy of big data (Hardjono et al., 2019). This will increase the likelihood of building and implementing scalable systems for collecting big data while putting more emphasis on relevant information to achieve user privacy.

8.2. Enforcement of secure transport and encryption mechanisms

Cloud computing technology made it possible for communication networks to connect many people around the world. The encryption of data-centric networks is essential for all users that use the network to communicate and share files (Choksi et al., 2015). To guarantee the security of data, only the authorized entities can be allowed to access control policies set for changing and saving the data. A uniform authentication agreement and fairness between the distributed network entities can be applied to ensure the communication framework is cryptographically secured or not.

Depending on the service provider, it might be possible that sensitive data is generally stored unencrypted in the cloud because some companies may find it expensive to plan for encryption services and hire experts and consultants in the field of cryptography. The all-or-nothing retrieval policy of encrypted data tends to restrict end-users from manipulating their data before storage and while being conveyed. Clients expect to share records or searches easily which needs better-performing networks. Companies that can afford good encryption algorithms often alleviate this problem by using cryptosystems such as public keys based on entity attributes used to encrypt sensitive data. Alternatively, data that might be useful for analytics but considered to be less sensitive and is, therefore, unencrypted ought to be communicated via a cryptographically secure communication framework.

8.3. Granular access control

The control and customization of user access are important as this defines the owner of the data and the set of permissions agreed upon for the user to manipulate the information. Security seems to be a lesser point of focus for companies that presently use Big Data solutions. This is because most of these applications are designed for performance and scalability. Compared to NoSQL DBs, the traditional relational databases have better security features (Alharthi et al., 2017). In Big Data, the terms of access control, the tables, columns, rows, cells, and authorized users are still a fundamental challenge. The main problem with coarse-grained access methods is that the big data that is meant to be shared is often redirected into more restrictive groups to enforce security which means the user does not have access at that particular time (Choksi et al., 2015).

Granular access control is necessary for Artificial Intelligence (AI) systems and analytical systems to adapt big data networks that are increasingly becoming complex to be secured due to the inherently diverse environment. As Choksi et al. (2015) opined, the costly nature of keeping a log of roles and building a list of authorized users is one of the biggest issues threatening big data firms from accomplishing analytical transformations. Lekshmy and Rahiman (2020) asserted that companies participating in data mining and big data analysis through cloud computing storage have to progressively focus on handling large and diverse data sets while ensuring that all the security restrictions are installed (See 4. above).

Embedding legal and policy restrictions on data communication networks for different data schemas and sources to match corporate and government privacy policies impose cost requirements on data handling. The cost might include management fees for the excess restrictions integrated into the developed applications and charges from the building as well as maintaining network security through a walled garden approach. This mechanism ensures there is comprehensive access control of the data centers with only a few approved people allowed to participate in collecting data and using it for positive or functional analysis.

9. Data management

Poor data management is a major reason for violating the privacy of big data, as users responsible for protecting customer data fail to implement security measures designed to protect data applications, especially those running on the company's network infrastructure. Data management activities include:

9.1. Secure data storage and transactions logs

The security problems with big data and its storage are related to the management and processing of the data. The lack of security in big data storage as well as the transaction logs may corrupt the data due to unlawful access by a malicious individual (See section 7.1 above). Eavesdropping big data during transferring from one source to another is a common cause for privacy violation and data tampering which interferes with the data delivery to the end-users (Choksi et al., 2015). Transaction logs on server computers tend to be stored in multi-tiered data storage media, thus, securing such data logs requires manual operations to a certain extent between the tiers. This gives network administrators and software application developers direct control over who and when they can access and move vital data from the data storage devices.

However, complex network systems that hold big data as it grows exponentially in volume require the use of auto-tiering for big data storage and management to guarantee the scalability and availability of the data. Nevertheless, auto-tiering tools and solutions cannot monitor the movement of data in transit or storage which begs the question of whether it is secure or not. For example, if an organization's internal cloud computing system wants to incorporate data from different departments, much of the data is never accessed or retrieved when it is in concurrent use in other divisions. Hence, an auto-tier data storage scheme would help the company save money by organizing the rarely utilized data to lower data pools. But, even if the data tier may contain less sensitive critical information, the organization ought to review the data arrangement strategies to ensure the lower data pools are also provided with security (Jain et al., 2016).

9.2. Granular audits

Prompt incidence response teams are imperative for companies when it comes to real-time security monitoring. Getting a notification of an actual or attempted attack at the moment it occurs is the objective of security systems (Anshari et al., 2019). However, in real life, this is not always the case because sometimes network audits have to be performed to determine missed attacks as well as set up new security measures for mitigating any future failures. The information collected during auditing processes is insightful in understanding the causes of privacy breaches on big data storage (Choksi et al., 2015). Extensive auditing and forensic investigation ensure that a company's resources and employees comply with regulations when handling big data trends within the organization and over distributed network processes. Therefore, auditing capabilities need to be executed across big data frameworks depending on the inspection tools and features enabled for the infrastructure modules (Jain et al., 2016). Examples of auditing and compliance include conducting Syslog (system log) on routers, software application logging as well as facilitating logging on the network end nodes operating systems.

9.3. Data provenance

Data provenance identifies and defines the attributes of the data origin or source. In its resilient form, big data that highlights provenance maintains the information and process integrity. This is accomplished through the documentation of systems IDs, the entities, and processes operating on precise data blocks of interest. Being able to keep track of data origins and document the historical changes applied to big data blocks helps keep a record of the data characteristics and state over the information's lifetime (Choksi et al., 2015). The exploration of data sources and attributes to detect security dependencies and confidentiality in software applications is computationally intensive. Therefore, these security assessments require fast algorithms to handle the provenance of big data blocks such as details about the date and time of its creation. An example is in cases that involve the assessment of financial systems that investigate insider trading for stock market companies.

10. Stakeholder role in big data security

The Big Data stakeholders include giant technology companies such as Google, Uber, Yahoo, and Facebook that form part of the leading firms in collecting client and user data. Social media platforms have been at the forefront of collecting all sorts of data inclusive of opinions, views, shopping preferences, and credit card data. YU (2016) reported that continuous mining and acquisition of data by big data companies are slowly lowering the security standards expected of data storage. These big companies ought to ensure the security of data stored in the databases by protecting it from different third parties or external entities that are not given access and authorization.

However, the initiative of big data for many companies can still be considered to be a profiling tool through which malicious individuals with access to individuals' private data can use it to intentionally stigmatize users. For example, this can be done through monitoring end-users social media activity and revealing their most embarrassing moments posted online. Yet, though big data analysis can be useful in analyzing people's preferences and likes to better customize one's shopping experience and interests, in big data, this is also considered to be an invasion of privacy. For instance, hacking and accessing social media accounts belonging to terrorists to avert attacks is also a privacy breach. Despite this disadvantage, the benefits of big data are much as it provides better off solutions to applications in the areas of healthcare, business, state governance (Remiche et al., 2019). Stakeholders ought to be encouraged to continuously contribute to the growth of big data while mitigating the existing concerns with big data such as privacy.

Some of the individual contributions that can be seen today that try to improve big data privacy can be seen in companies that take precedence in protecting their private data. For example, two of the biggest names in technology recently got into a legal scuffle regarding self-driving vehicles, Google's parent company, Alphabet, sued Uber for stealing its intellectual information to develop self-driving vehicles using their technology. Before the legal situation was exacerbated, Uber had to resolve the case out of court paying Google over 200 million dollars in legal settlement fees. Moreover, the culprit responsible for stealing sensitive information about Google's driverless car and sharing the data was a former Google employee, who was later sentenced to do prison time. Competition is healthy for business, although using hand-to-hand tactics to compromise client or company data is known to set bad precedence between business competitors.

The Big Companies that deal with volumes of data are always in search of better opportunities and growing the business via attracting prospective clients. The laws introduced to govern big data are currently in the growth stage, hence, despite the effort to make changes in various IT departments in major technology companies, guaranteeing security is still a big challenge. However, big data companies might not care about the reputational damage that privacy breaches may cause because some of these companies are willing to see their visions and goals achieved at all costs. This means that big data stakeholders with an interest in mining customer information and selling to third parties will do so despite imminent lawsuits for data breaches as long as the outcome is profitable to the company. Just like Uber, some big companies are willing and confident that they will triumph over their court cases and still earn revenue from the trade secrets acquired via data breaches. It is clear that Big Data presents interesting opportunities for individual users and companies; however, these chances are contradicted by big challenges in terms of security and privacy (Pape and Stankovic, 2019). The traditional security mechanisms that are aimed at delivering awareness to people are insufficient to offer a proficient answer to those challenges. In the next section, some of these solutions and proposals are going to be addressed. These include initiatives such as training big data firms' staff and clients on reactive security awareness and mitigation approaches that can be utilized to respond to past and current security threats.

11. Proposed security methods

There is no single enhanced approach to solve the issue of Big Data privacy breaches and security challenges. The current traditional security mechanisms that are designed to secure small-scale business entities are frequently used to secure static data; however, these security tools often fall short when it comes to protecting big data (Remiche et al., 2019). Data experts and common computer users must comprehend how the collection of big data can be protected whether the data are structured or unstructured. Big companies can have unauthorized access to client data to create new relations; hence, combining different data elements allows malicious users to exploit the Big Data. The essential solution for Big Data privacy challenges is the encryption of everything to make the client data safe regardless of the state whether it is transit or storage on a basic home computer, mobile device, a cloud-point server, or any data storage gadget. The encryption, masking, and tokenization of sensitive information are the recommended approaches for protecting critical data.

Due to the complex features, Big Data projects require an all-around approach in implementing security. Big Data projects often have to take into consideration the documentation of the diverse data sources, the source, and creators of data, and the users permitted to access and manipulate the data. Additionally, experts have to perform the appropriate sorting to mark critical data to ensure the data is stored in a manner that complies with the governments' and companies' information security policy. As a recommendation, adaptive security methods such as encryption services ought to be installed on the network protocols and on the big data itself, to deliver better security for handling data at the source. Other proposed mechanisms of control and prevention in the big data storage area include a mix of access control and data leakage prevention strategies that work together to protect the big data (Govindarajan, 2019).

The new Big Data security answers should entail the securing perimeter walls of a business organization as well as the technology devices such as the public cloud computing environment. Likewise, a reliable data provenance method ought to be developed to support different domains for a single business entity. Moreover, similar security mechanisms such as the ones highlighted by Govindarajan (2019) can be utilized to alleviate cyberattacks like a distributed denial-of-service (DDoS) attacks targeted against Big Data set-ups. Also, in the whole lifecycle of Big Data, security and privacy play a vital role in ensuring that the information's trustworthiness remains the same from data collection to usage. The customization characteristics of certain Big Data services and their effect on user privacy are conferred in Govindarajan (2019) where the authors discuss these challenges in the backdrop of EEXCESS. This is a data security project that ventures in both delivering high-level recommendations for user privacy. A recent study emphasizes privacy extensions to Unified Modelling Language (UML) to help software developers visualize demonstrate privacy requirements. It helps software engineers to subsequently incorporate privacy needs into the Software Life Development Cycle (SDLC) of Big Data applications (Martinez-Mosquera et al., 2019).

Big Data security mandates the tools that address legal necessities about data handling must be met to ensure security and privacy are maintained. Secure encryption technology ought to be instigated to protect all the confidential data that is related to a computer user. The sensitive information includes Intellectual Property (IP), Personally Identifiable Information (PII), and Protected Health Information (PHI) (Viceconti et al., 2015). Security procedures need to put in place using cryptography and access management policies to ensure the accurate encoding and decoding of data in stores or being conveyed via the internet. All these mechanisms inclusive of software and hardware-based encryptions are expected to be transparent to the end-user while guaranteeing that performance and scalability of data networks of systems do not deteriorate (Govindarajan, 2019). As previously discussed, traditional anonymization and encryption of data are not enough to solve Big Data privacy issues as they are only sufficient to protect static data and not a computational analysis of data held by small businesses (Hardjono et al., 2019).

Consequently, other methods that allow for precise and targeted data to be analyzed while keeping the data encrypted ought to be utilized. For instance, techniques such as Fully Homomorphic Encryption (FHE) (El-Yahyaoui and El Kettani, 2017), Secure Function Evaluation (SFE) (Lekshmy and Rahiman, 2020), and Functional Encryption (FE) (Goodrich et al., 2014) are advisable for implementation. Moreover, the partitioning of data storage devices on non-communicating data centers can aid in lessening the problems that limit the execution of traditional security techniques. Homomorphic data encryption is a type of cryptography that allows specific types of computations. For instance, RSA public key algorithm is a protective approach that can be implemented on ciphertext to generate encryption and decryption codes that match the results translated from encryption operations performed on plaintext (YU, 2016).

Fully homomorphic encryption is a security solution that delivers numerous applications including cloud computing platforms for big data, as mentioned in El-Yahyaoui and El Kettani (2017). The solution enables database administrators to perform encrypted queries on databases, which ensures that user data is always kept private in the data stores. Common data stores used by big data entail cloud storage either internally or on devices provided by suppliers. Cloud service providers sometimes provide servers that do not encrypt the data stored on them. Hence, this limits any user who stores data on these untrusted servers but if a client encrypts their information before storage on such servers they do not have to worry about the data secrecy (Nasereddin and Darwesh, 2020). Encrypted queries also help users conduct private searches to search engine forms. For example, when a user submits a search query, the search engine scrambles the data as well as computes a concise encrypted answer without clearly revealing the query details which could have sensitive user information such as their tax filing number or even their private health records.

Homomorphic encryption services also facilitate the searching of encrypted data stored on a database. For instance, end-users that store files on a remote file server in an encrypted format can have the server recover only files that satisfy a certain Boolean constraint when the file is decoded. Moreover, fully homomorphic encryption improves the effectiveness of secure dual computation which means that even if the data storage servers cannot decrypt a file on their own, the service simplifies securing the communication channel. A significant Big Data privacy and security challenge for Big Data is associated with the storage and processing of encrypted data. Thus, being able to run queries against an encrypted database is a fundamental security requirement to protect Big Data however it is a challenging one. It introduces problems such as whether the datastore is encrypted with single or multiple keys. It also questions if the database ought to be decrypted before running the query; do the search queries have to be encrypted; and which people are granted permission to decrypt the database.

Recently, a security system called CryptDB developed at MIT attempts to provide an effective solution to some of these problems. CryptDB lets scientists run queries on records over encrypted data, which means they only retrieve what is important to the search query (Nasereddin and Darwesh, 2020). Accordingly, dependable and reliable applications that plan to query encrypted data must pass those queries via the CryptDB proxy (that monitors connections between the application and the database) then it subsequently makes alterations to those queries in an explicit manner so that they can be executed against the encrypted database. The concept behind the security system is that it uses the CryptDB proxy to hold master keys that are used to decrypt the search query results and send the absolute outcome back to the application. According to Nasereddin and Darwesh (2020); the CryptDB application supports many forms of data encryption patterns that permit different types of actions to be performed on the data A good example is the Encrypted Big Query Client developed and designed by Google Company to protect their Big Data. The Encrypted Big Query Client is based on similar operation models as CryptDB which allows encrypted big queries to be executed against their BigQuery service that enables super, SQL-like queries against append-only tables, utilizing the exponential processing power of Google's network infrastructure.

Apart from the defined security recommendations, it is vital to consider the security of the IT infrastructure itself. The best security practices to put into place are security measures that control the network's perimeter and peripheral devices that store sensitive data. Nevertheless, if an attacker exploits and violates the security perimeter, it might allow them to gain access to all the data within the set IT frameworks. Hence, a new method is essential to migrate those security controls close to the data centers. Leverage the already existing security procedures such a data tracking, analyzing, and learning from data utilization and access is also a key feature to constantly advance the security of the computing servers that hold sensitive data (Alharthi et al., 2017).

12. Conclusion

As Big Data expands and grows on public cloud infrastructure so does the need to have control over the cloud infrastructure. Through a proper examination of both static and streaming large data sets, companies will be able to make better improvements in numerous scientific fields and medical disciplines. Traditional security methodologies often address small-scale business systems that hold static data on semi-isolated networks, while, securing Big Data necessitates the use of different approaches to ensure profitability for many big companies. Threats to Big Data privacy increase exponentially when enterprise users lose chain-of-custody control of their data or become dependent on closed, enterprise systems that require consumers to surrender their data to vendors or merchants. That's why it is vital to select Big Data frameworks that are open, permit commercial clients and government agencies to approve who can see aggregated but private data. It is practically improbable to envision the next generation of computer applications without using and storing data via data-driven algorithms.

The complexities presented by the compression of data, encryption, access control, and compliance ought to be addressed in a methodical way. This guarantees security as computing infrastructure becomes affordable, software applications and operating systems become networked, and data analytics settings are shared over the cloud computing platforms. This paper has investigated and highlighted the top privacy and security difficulties that must be addressed to make Big Data operations and computing infrastructure more secure. It is the belief that this paper will inspire action in the research and development of society to collectively focus on the obstacles to higher security and privacy in Big Data.

In this paper, we introduced the overview of some of the modern methods that can be utilized suggested in the field of big data such as:

1. Data governance–the paper analyzes common data representation and the methods used to secure Big Data according to industry standards and local and regional regulations.

2. Real-time security analytics-analyzing security risks and forecasting threat sources in real-time is of utmost need in the big data industry. The use of real-time security systems such as anti-malware help Big Data companies secure data stored on their devices.
3. Privacy preservation-analytics that is based on preserving user privacy in the domain of big data analytics. Privacy-preserving encryption methods allow running prediction algorithms on encrypted data to come up with reliable outputs while protecting the user data.
4. Using predictive models in data mining-to ensure faster analysis of Big Data while guaranteeing that the models used are industry standard.
5. Ensuring quality privacy–guaranteeing the quality of the data through ensuring the privacy of big data sharing methods used by the organizations.

Acknowledgment

The authors would like to acknowledge the support provided by AlMaarefa University while conducting this research work.

Compliance with ethical standards

Conflict of interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

Alharthi A, Krotov V, and Bowman M (2017). Addressing barriers to big data. Business Horizons, 60(3): 285-292. https://doi.org/10.1016/j.bushor.2017.01.002

Anshari M, Almunawar MN, Lim SA, and Al-Mudimigh A (2019). Customer relationship management and big data enabled: Personalization and customization of services. Applied Computing and Informatics, 15(2): 94-101. https://doi.org/10.1016/j.aci.2018.05.004

Atoum I and Al-Jarallah NA (2019). Big data analytics for value-based care: Challenges and opportunities. International Journal of Advanced Trends in Computer Science and Engineering, 8(6): 3012-3016. https://doi.org/10.30534/ijatcse/2019/55862019

Choksi K, Dalal N, Gupte MK, and Jivani A (2015). Security and privacy challenges in big data. International Journal of Latest Trends in Engineering and Technology, 7(3): 313-318. https://doi.org/10.21172/1.73.543

Davenport T, Guha A, Grewal D, and Bressgott T (2020). How artificial intelligence will change the future of marketing. Journal of the Academy of Marketing Science, 48(1): 24-42. https://doi.org/10.1007/s11747-019-00696-0

El-Yahyaoui A and El Kettani MD (2017). A verifiable fully homomorphic encryption scheme to secure big data in cloud computing. In the International Conference on Wireless Networks and Mobile Communications (WINCOM), IEEE, Rabat, Morocco: 1-5. https://doi.org/10.1109/WINCOM.2017.8238186

Goodrich MT, Tamassia R, and Goldwasser MH (2014). Data structures and algorithms in Java. John Wiley and Sons, Hoboken, USA.

Govindarajan M (2019). Challenges for big data security and privacy. In: Khosrow-Pour DBAM (Ed.), Advanced Methodologies and Technologies in Network Architecture, Mobile Computing, and Data Analytics: 57-66. http://doi:10.4018/978-1-5225-7598-6.ch005

Hardjono T, Shrier DL, and Pentland A (2019). Trusted data: A new framework for identity and data sharing. MIT Connection Science and Engineering, Cambridge, USA. https://doi.org/10.7551/mitpress/12439.001.0001

Jain P, Gyanchandani M, and Khare N (2016). Big data privacy: A technological perspective and review. Journal of Big Data, 3: 25. https://doi.org/10.1186/s40537-016-0059-y

Lekshmy PL and Rahiman MA (2020). A sanitization approach for privacy preserving data mining on social distributed environment. Journal of Ambient Intelligence and Humanized Computing, 11: 2761–2777. https://doi.org/10.1007/s12652-019-01335-w

Martin KD, Borah A, and Palmatier RW (2017). Data privacy: Effects on customer and firm performance. Journal of Marketing, 81(1): 36-58. https://doi.org/10.1509/jm.15.0497

Martinez-Mosquera D, Luján-Mora S, Navarrete R, Mayorga TC, and Vivanco Herrera HR (2019). An approach to big data modeling for key-value NoSQL databases. Iberian Journal of Information Systems and Technologies, E19: 519-530.

Moura J and Serrão C (2015). Security and privacy issues of big data. In: Zaman N, Seliaman ME, Hassan MF, and Márquez FPG (Eds.), Handbook of research on trends and future directions in big data and web intelligence: 20-52. IGI Global, Pennsylvania, USA.

Nasereddin HH and Darwesh AJ (2020). An object oriented programming on encrypted database system (CryptDB). Journal of Talent Development and Excellence, 12(1): 5140-5146.

Pape S and Stankovic J (2019). An insight into decisive factors in cloud provider selection with a focus on security. In: Katsikas S, Cuppens F, Cuppens N, Lambrinoudakis C, Kalloniatis C, Garcia-Alfaro J (Eds.), Computer security: 287-306. Springer, Cham, Switzerland.

Perera C, Ranjan R, Wang L, Khan SU, and Zomaya AY (2015). Big data privacy in the internet of things era. IT Professional, 17(3): 32-39. https://doi.org/10.1109/MITP.2015.34

Reis T, Bornschlegl MX, and Hemmje ML (2020). Big data analysis, AI, and visualization workshop: Road mapping infrastructures for artificial intelligence supporting advanced visual big data analysis. In the International Conference on Advanced Visual Interfaces, Association for Computing Machinery, Salerno, Italy: 1-2. https://doi.org/10.1145/3399715.3400860

Remiche JM, Aubert J, Mayer N, and Petrocelli D (2019). Evaluation of cloud computing offers through security risks. In the 8th International Conference on Cloud Computing and Services Science, Funchal, Madeira Island, Portugal: 338-345. https://doi.org/10.5220/0006665703380345

Sardi A, Sorano E, Cantino V, and Garengo P (2020). Big data and performance measurement research: Trends, evolution and future opportunities. Measuring Business Excellence. https://doi.org/10.1108/MBE-06-2019-0053

Singh C (2020). Phishing website detection based on machine learning: A survey. In the 6th International Conference on Advanced Computing and Communication Systems, IEEE, Coimbatore, India: 398-404. https://doi.org/10.1109/ICACCS48705.2020.9074400

Suoniemi S, Meyer-Waarden L, Munzel A, Zablah AR, and Straub D (2020). Big data and firm performance: The roles of market-directed capabilities and business strategy. Information and Management, 57(7): 103365. https://doi.org/10.1016/j.im.2020.103365

Viceconti M, Hunter P, and Hose R (2015). Big data, big knowledge: Big data for personalized healthcare. IEEE Journal of Biomedical and Health Informatics, 19(4): 1209-1215. https://doi.org/10.1109/JBHI.2015.2406883 **PMid:26218867**

Wu Y, Huang H, Wu N, Wang Y, Bhuiyan MZA, and Wang T (2020). An incentive-based protection and recovery strategy for secure big data in social networks. Information Sciences, 508: 79-91. https://doi.org/10.1016/j.ins.2019.08.064

Yu S (2016). Big privacy: Challenges and opportunities of privacy study in the age of big data. IEEE Access, 4: 2751-2763. https://doi.org/10.1109/ACCESS.2016.2577036