Abouelmehdi *et al. J Big Data (2018) 5:1*
https://doi.org/10.1186/s40537-017-0110-7

Journal of Big Data

SURVEY PAPER Open Access

Big healthcare data: preserving security and privacy

Karim Abouelmehdi*, Abderrahim Beni-Hessane and Hayat Khaloufi

*Correspondence: karim.abouelmehdi1@gmail.com
Department of Computer Science Laboratory LAMAPI and LAROSERI, Chouaib Doukkali University, El Jadida, Morocco

Abstract

Big data has fundamentally changed the way organizations manage, analyze and leverage data in any industry. One of the most promising fields where big data can be applied to make a change is healthcare. Big healthcare data has considerable potential to improve patient outcomes, predict outbreaks of epidemics, gain valuable insights, avoid preventable diseases, reduce the cost of healthcare delivery and improve the quality of life in general. However, deciding on the allowable uses of data while preserving security and patient's right to privacy is a difficult task. Big data, no matter how useful for the advancement of medical science and vital to the success of all healthcare organizations, can only be used if security and privacy issues are addressed. To ensure a secure and trustworthy big data environment, it is essential to identify the limitations of existing solutions and envision directions for future research. In this paper, we have surveyed the state-of-the-art security and privacy challenges in big data as applied to healthcare industry, assessed how security and privacy issues occur in case of big healthcare data and discussed ways in which they may be addressed. We mainly focused on the recently proposed methods based on anonymization and encryption, compared their strengths and limitations, and envisioned future research directions.

Keywords: Security and privacy, Big healthcare data, Security lifecycle, Anonymization, Encryption

Introduction

Change is the new norm for the global healthcare sector. In fact, digitization of health and patient data is undergoing a dramatic and fundamental shift in the clinical, operating and business models and generally in the world of economy for the foreseeable future. This shift is being spurred by aging populations and lifestyle changes; the proliferation of software applications and mobile devices; innovative treatments; heightened focus on care quality and value; and evidence-based medicine as opposed to subjective clinical decisions—all of which are leading to offer significant opportunities for supporting clinical decision, improving healthcare delivery, management and policy making, surveilling disease, monitoring adverse events, and optimizing treatment for diseases affecting multiple organ systems [1, 2].

As noted above, big data analytics in healthcare carries many benefits, promises and presents great potential for transforming healthcare, yet it raises manifold barriers and challenges. Indeed, the concerns over the big healthcare data security and privacy are increased year-by-year. Additionally, healthcare organizations found that a reactive,

bottom-up, technology-centric approach to determining security and privacy requirements is not adequate to protect the organization and its patients [3].

Motivated thus, new information systems and approaches are needed to prevent breaches of sensitive information and other types of security incidents so as to make effective use of the big healthcare data.

In this paper, we discuss some interesting related works and present risks to the big health data security as well as some newer technologies to redress these risks. Then, we focus on the big data privacy issue in healthcare, by mentioning various laws and regulations established by different regulatory bodies and pointing out some feasible techniques used to ensure the patient's privacy. Thereafter, we provide some proposed techniques and approaches that were reported in the literature to deal with security and privacy risks in healthcare while identifying their limitations. Lastly, we offer conclusions and highlight the future directions.

Successful related works

Seamless integration of greatly diverse big healthcare data technologies can not only enable us to gain deeper insights into the clinical and organizational processes but also facilitate faster and safer throughput of patients and create greater efficiencies and help improve patient flow, safety, quality of care and the overall patient experience no matter how costly it is.

Such was the case with South Tyneside NHS Foundation Trust, a provider of acute and community health services in northeast England that understands the importance of providing high quality, safe and compassionate care for the patients at all times, but needs a better understanding of how its hospitals operate to improve resource allocation and wait times and to ensure that any issues are identified early and acted upon [4].

Another example is the UNC Health Care (UNCHC), which is a non-profit integrated healthcare system in North Carolina that has implemented a new system allowing clinicians to rapidly access and analyze unstructured patient data using natural-language processing. In fact, UNCHC has accessed and analyzed huge quantities of unstructured content contained in patient medical records to extract insights and predictors of readmission risk for timely intervention, providing safer care for high-risk patients and reducing re-admissions [5].

Moreover in the United States, the Indiana Health Information Exchange, which is a non-profit organization, provides a secure and robust technology network of health information linking more than 90 hospitals, community health clinics, rehabilitation centers and other healthcare providers in Indiana. It allows medical information to follow the patient hosted in one doctor office or only in a hospital system [6].

One more example is Kaiser Permanente medical network based in California. It has more than 9 million members, estimated to manage large volumes of data ranging from 26.5 Petabytes to 44 Petabytes. [7].

Big data analytics is used also in Canada, e.g. the infant hospital of Toronto. This hospital succeeded to improve the outcomes for newborns prone to serious hospital infections. Another example is the Artemis project, which is a newborns monitoring platform designed mercy to a collaboration between IBM and the Institute of Technology of Ontario. It supported the acquisition and the storage of patients' physiological data and

clinical information system data for the objective of online and real time analysis, retrospective analysis, and data mining [8].

In Europe and exactly in Italy, the Italian medicines agency collects and analyzes a large amount of clinical data concerning expensive new medicines as part of a national profitability program. Based on the results, it may reassess the medicines prices and market access terms [9].

In the domain of mHealth, the World Health Organization has launched the project "Be Healthy Be mobile" in Senegal and under the mDiabetes initiative it supports countries to set up large-scale projects that use mobile technology, in particular text messaging and apps, to control, prevent and manage non-communicable diseases such as diabetes, cancer and heart disease [10]. mDiabetes is the first initiative to take advantage of the widespread mobile technology to reach millions of Senegalese people with health information and expand access to expertise and care. Launched in 2013, in Costa Rica that has been officially selected as the first country, the initiative is working on an mCessation for tobacco program for smoking prevention and helping smokers quit, an mCervical cancer program in Zambia and has plans to roll out mHypertension and mWellness programs in other countries.

After Europe, Canada, Australia, Russia, and Latin America, Sophia Genetics [11], global leader in data-driven medicine, announced at the recent 2017 Annual Meeting of the American College of Medical Genetics and Genomics (ACMG) that its artificial intelligence has been adopted by African hospitals to advance patient care across the continent.

In Morocco for instance, PharmaProcess in Casablanca, ImmCell, The Al Azhar Oncology Center and The Riad Biology Center in Rabat are some medical institutions at the forefront of innovation that have started integrating Sophia to speed and analyze genomic data to identify disease-causing mutations in patients' genomic profiles, and decide on the most effective care. As new users of SOPHIA, they become part of a larger network of 260 hospitals in 46 countries that share clinical insights across patient cases and patient populations, which feeds a knowledge-base of biomedical findings to accelerate diagnostics and care [12].

While the automations have led to improve patient care workflow and reduce costs, it is also rising healthcare data to increase probability of security and privacy breaches. In 2016, CynergisTek has released the Redspin's 7th annual breach report: Protected Health Information (PHI) [13] in which it has reported that hacking attacks on healthcare providers were increased 320% in 2016, and that 81% of records breached in 2016 resulted from hacking attacks specifically. Additionally, ransomware, defined as a type of malware that encrypts data and holds it hostage until a ransom demand is met, has identified as the most prominent threat to hospitals. Additional findings of this report include:

- 325 large breaches of PHI, compromising 16,612,985 individual patient records.
- 3,620,000 breached patient records in the year's single largest incident.
- 40% of large breach incidents involved unauthorized access/disclosure.

These findings point to a pressing need for providers to take a much more proactive and comprehensive approach to protecting their information assets and combating the growing threat that cyber attacks present to healthcare.

Several prosperous initiatives have appeared to help the healthcare industry continually improve its ability to protect patient information.

In January 2014, for example, the White House, led by President Obama's Counselor John Podesta, undertook a 90-day review of big data and privacy. The review brought concrete recommendations to maximize benefits and minimize risks of big data [14, 15], namely:

- Policy attention should focus more on the actual uses of big data and less on its collection and analysis. Such existing policies are unlikely to yield effective strategies for improving privacy, or to be scalable over time.
- Policy concerning privacy protection should be addressing the purpose rather than prescribing the mechanism.
- Research is needed in the technologies that help to protect privacy, in the social mechanisms that influence privacy preserving behavior, and in the legal options that are robust to changes in technology and create appropriate balance among economic opportunity, national priorities, and privacy protection.
- Increased education and training opportunities concerning privacy protection, including career paths for professionals. Programs that provide education leading to privacy expertise are essential and need encouragement.
- Privacy protections should be extended to non-US citizens as privacy is a worldwide value that should be reflected in how the federal government handles personally identifiable information from non-US citizens [16].

The OECD Health Care Quality Indicators (HCQI) project is responsible for a plan in 2013/2014 to develop tools to assist countries in balancing data privacy risks and risks from not developing and using health data. This plan includes developing a risk categorization of different types and uses of data and the promising practices that countries can deploy to reduce risks that directly affect everyone's daily life and enable data use [17].

Privacy and security concerns in big data

Security and privacy in big data are important issues. Privacy is often defined as having the ability to protect sensitive information about personally identifiable health care information. It focuses on the use and governance of individual's personal data like making policies and establishing authorization requirements to ensure that patients' personal information is being collected, shared and utilized in right ways. While security is typically defined as the protection against unauthorized access, with some including explicit mention of integrity and availability. It focuses on protecting data from pernicious attacks and stealing data for profit. Although security is vital for protecting data but it's insufficient for addressing privacy. Table 1 focuses on additional difference between security and privacy.

Table 1 Differentiation between security and privacy

Security	Privacy
Security is the "confidentiality, integrity and availability" of data	Privacy is the appropriate use of user's information
Various techniques like Encryption, Firewall, etc. are used in order to prevent data compromise from technology or vulnerabilities in the network of an organization	The organization can't sell its patient/user's information to a third party without prior consent of the user
It may provide for confidentiality or protect an enterprise or agency	It concerns with patient's right to safeguard their information from any other parties
Security offers the ability to be confident that decisions are respected	Privacy is the ability to decide what information of an individual goes and where to

Security of big healthcare data

While healthcare organizations store, maintain and transmit huge amounts of data to support the delivery of efficient and proper care, the downsides are the lack of technical support and minimal security. Complicating matters, the healthcare industry continues to be one of the most susceptible to publicly disclosed data breaches. In fact, attackers can use data mining methods and procedures to find out sensitive data and release it to the public and thus data breach happens. Whereas implementing security measures remains a complex process, the stakes are continually raised as the ways to defeat security controls become more sophisticated.

Accordingly, it is critical that organizations implement healthcare data security solutions that will protect important assets while also satisfying healthcare compliance mandates.

A. Big data security lifecycle

In terms of security and privacy perspective, Kim et al. [18] argue that security in big data refers to three matters: data security, access control, and information security. In this regards, healthcare organizations must implement security measures and approaches to protect their big data, associated hardware and software, and both clinical and administrative information from internal and external risks. At a project's inception, the data lifecycle must be established to ensure that appropriate decisions are made about retention, cost effectiveness, reuse and auditing of historical or new data [19].

Yazan et al. [20] suggested a big data security lifecycle model extended from Xu et al. [21]. This model is designed to address the phases of the big data lifecycle and correlate threats and attacks that face big data environment within these phases, while [21] address big data lifecycle from user role perspective: data provider, data collector, data miner, and decision maker. The model proposed in [20] comprised of four interconnecting phases: data collection phase, data storage phase, data processing and analysis, and knowledge creation.

Furthermore, CCW (The Chronic Conditions Data Warehouse) follows a formal information security lifecycle model, which consists of four core phases that serve to identify, assess, protect and monitor against patient data security threats. This lifecycle model is continually being improved with emphasis on constant attention and continual monitoring [21].

In this paper, we suggest a model that combines the phases presented in [20] and phases mentioned in [21], in order to provide encompass policies and mechanisms that ensure addressing threats and attacks in each step of big data life cycle. Figure 1 presents the main elements in big data lifecycle in healthcare.

- *Data collection phase* This is the obvious first step. It involves collecting data from different sources in various formats. From a security perspective, securing big health data technology is a necessary requirement from the first phase of the lifecycle. Therefore, it is important to gather data from trusted sources, preserve patient privacy (there must be no attempt to identify the individual patients in the database) and make sure that this phase is secured and protected. Indeed, some mature security measures must be used to ensure that all data and information systems are protected from unauthorized access, disclosure, modification, duplication, diversion, destruction, loss, misuse or theft.
- *Data transformation phase* Once the data is available, the first step is to filter and classify the data based on their structure and do any necessary transformations in order to perform meaningful analysis. More broadly, data filtering, enrichment and transformation are needed to improve the quality of the data ahead of analytics or modeling phase and remove or appropriately deal with noise, outliers, missing values, duplicate data instances, etc. On the other side, the collected data may contain

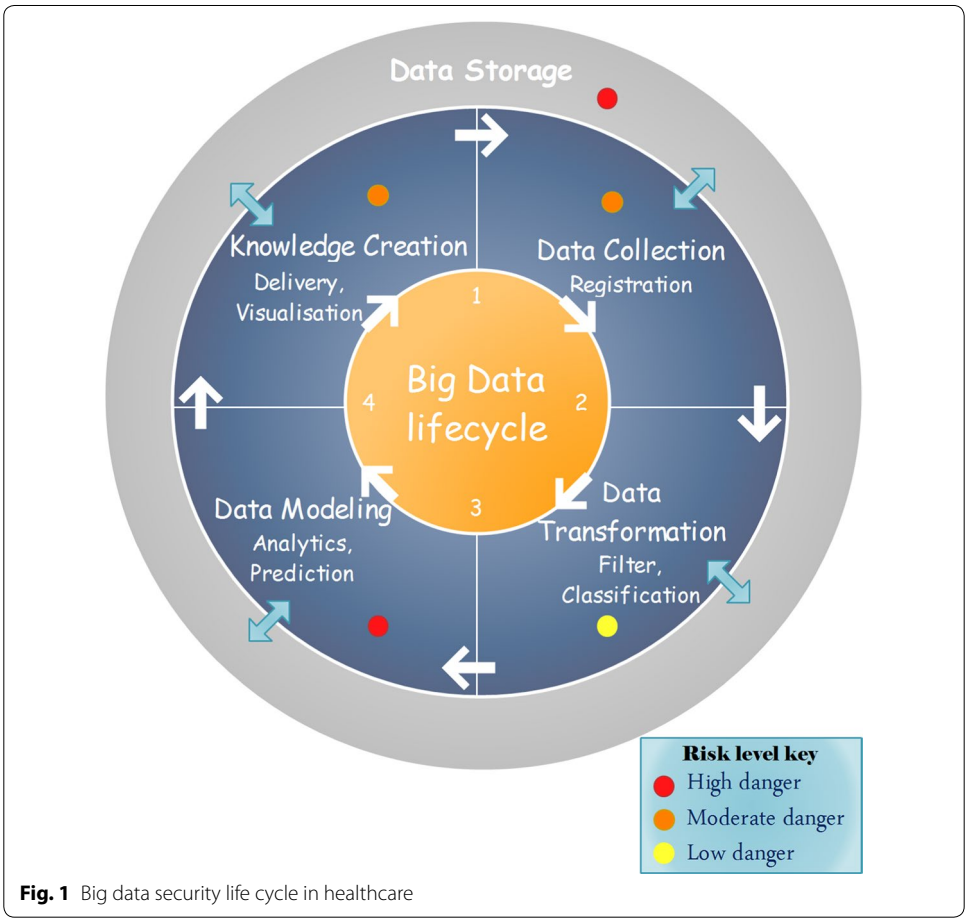

Fig. 1 Big data security life cycle in healthcare

sensitive information, which makes extremely important to take sufficient precautions during data transformation and storing. In order to guarantee the safety of the collected data, the data should remain isolated and protected by maintaining access-level security and access control (utilizing an extensive list of directories and databases as a central repository for user credentials, application logon templates, password policies and client settings) [22], and defining some security measures like data anonymization approach, permutation, and data partitioning.

- *Data modeling phase* Once the data has been collected, transformed and stored in secured storage solutions, the data processing analysis is performed to generate useful knowledge. In this phase, supervised data mining techniques such as clustering, classification, and association can be employed for feature selection and predictive modeling. Further, there also exist several ensembles of learning techniques that improve accuracy and robustness of the final model. On the other side, it is crucial to provide secure processing environment. In fact, the focus of data miners in this phase is to use powerful data mining algorithms that can extract sensitive data. Therefore, the process of data mining and the network components in general, must be configured and protected against data mining based attacks and any security breach that may happen, as well as make sure that only authorized staff work in this phase. This process helps eliminate some vulnerabilities and mitigates others to a lower risk level.
- *Knowledge creation phase* Finally, the modeling phase comes up with new information and valued knowledges to be used by decision makers. These created knowledges are considered sensitive data, especially in a competitive environment. Indeed, healthcare organizations aware of their sensitive data (e.g. patient personal data) not to be publicly released. Accordingly, security compliance and verification are a primary objective in this phase.

At all stages of big data lifecycle, it requires data storage, data integrity and data access control.

B. Technologies in use

Various technologies are in use to ensure security and privacy of big healthcare data. Most widely used technologies are:

1) Authentication Authentication is the act of establishing or confirming claims made by or about the subject are true and authentic. It serves vital functions within any organization: securing access to corporate networks, protecting the identities of users, and ensuring that the user is really who he is pretending to be.

The information authentication can pose special problems, especially man-in-the-middle (MITM) attacks. Most cryptographic protocols include some form of endpoint authentication specifically to prevent MITM attacks. For instance [23], transport layer security (TLS) and its predecessor, secure sockets layer (SSL), are cryptographic protocols that provide security for communications over networks such as the Internet. TLS and SSL encrypt the segments of network connections at the transport layer end-to-end. Several versions of the protocols are in widespread use in applications like web browsing, electronic mail, Internet faxing, instant messaging and voice-over-IP (VoIP). One can use SSL or TLS to authenticate the server using a mutually trusted certification

authority. Hashing techniques like SHA-256 [24] and Kerberos mechanism based on Ticket Granting Ticket or Service Ticket can be also implemented to achieve authentication. Additionally, Bull Eye algorithm can be used for monitoring all sensitive information in 360°. This algorithm has been used to make sure data security and manage relations between original data and replicated data. It is also allowed only to an authorized person to read or write critical data. Paper [25] proposes a novel and simple authentication model using one time pad algorithm. It provides removing the communication of passwords between the servers. In a healthcare system, both healthcare information offered by providers and identities of consumers should be verified at the entry of every access.

2) Encryption Data encryption is an efficient means of preventing unauthorized access of sensitive data. Its solutions protect and maintain ownership of data throughout its lifecycle—from the data center to the endpoint (including mobile devices used by physicians, clinicians, and administrators) and into the cloud. Encryption is useful to avoid exposure to breaches such as packet sniffing and theft of storage devices.

Healthcare organizations or providers must ensure that encryption scheme is efficient, easy to use by both patients and healthcare professionals, and easily extensible to include new electronic health records. Furthermore, the number of keys hold by each party should be minimized.

Although various encryption algorithms have been developed and deployed relatively well (RSA, Rijndael, AES and RC6 [24, 26, 27], DES, 3DES, RC4 [28], IDEA, Blowfish …), the proper selection of suitable encryption algorithms to enforce secure storage remains a difficult problem.

3) Data masking Masking replaces sensitive data elements with an unidentifiable value. It is not truly an encryption technique so the original value cannot be returned from the masked value. It uses a strategy of de-identifying data sets or masking personal identifiers such as name, social security number and suppressing or generalizing quasi-identifiers like date-of-birth and zip-codes. Thus, data masking is one of the most popular approach to live data anonymization. k-anonymity first proposed by Swaney and Samrati [29, 30] protects against identity disclosure but failed to protect against attribute disclosure. Truta et al. [31] have presented p-sensitive anonymity that protects against both identity and attribute disclosure. Other anonymization methods fall into the classes of adding noise to the data, swapping cells within columns and replacing groups of k records with k copies of a single representative. These methods have a common problem of difficulty in anonymizing high dimensional data sets [32, 33].

A significant benefit of this technique is that the cost of securing a big data deployment is reduced. As secure data is migrated from a secure source into the platform, masking reduces the need for applying additional security controls on that data while it resides in the platform.

4) Access control Once authenticated, the users can enter an information system but their access will still be governed by an access control policy which is typically based on privileges and rights of each practitioner authorized by patient or a trusted third party. It is then, a powerful and flexible mechanism to grant permissions for users. It provides sophisticated authorization controls to ensure that users can perform only the activities

for which they have permissions, such as data access, job submission, cluster administration, etc.

A number of solutions have been proposed to address the security and access control concerns. Role-based access control (RBAC) [34] and attribute-based access control (ABAC) [35, 36] are the most popular models for EHR. RBAC and ABAC have shown some limitations when they are used alone in medical system. Paper [37] proposes also a cloud-oriented storage efficient dynamic access control scheme ciphertext based on the CP-ABE and a symmetric encryption algorithm (such as AES). To satisfy requirements of fine-grained access control yet security and privacy preserving, we suggest adopting technologies in conjunction with other security techniques, e.g. encryption, and access control methods.

5) Monitoring and auditing Security monitoring is gathering and investigating network events to catch the intrusions. Audit means recording user activities of the healthcare system in chronological order, such as maintaining a log of every access to and modification of data. These are two optional security metrics to measure and ensure the safety of a healthcare system [38].

Intrusion detection and prevention procedures on the whole network traffic is quite tricky. To address this problem, a security monitoring architecture has been developed via analyzing DNS traffic, IP flow records, HTTP traffic and honeypot data [39]. The suggested solution includes storing and processing data in distributed sources through data correlation schemes. At this stage, three likelihood metrics have been calculated to identify whether domain name, packet or flow is malicious. Depending on the score obtained through this calculation, an alert occurs in detection system or process terminate by prevention system. According to performance analysis with open source big data platforms on electronic payment activities of a company data, Spark and Shark produce fast and steady results than Hadoop, Hive and Pig [40].

Big data network security systems should be find abnormalities quickly and identify correct alerts from heterogeneous data. Therefore, a big data security event monitoring system model has been proposed which consists of four modules: data collection, integration, analysis, and interpretation [41]. Data collection includes security and network devices logs and event information. Data integration process is performed by data filtering and classifying. In data analysis module, correlations and association rules are determined to catch events. Finally, data interpretation provides visual and statistical outputs to knowledge database that makes decisions, predicts network behavior and responses events.

Privacy of big healthcare data

The invasion of patient privacy is considered as a growing concern in the domain of big data analytics due to the emergence of advanced persistent threats and targeted attacks against information systems. As a result, organizations are in challenge to address these different complementary and critical issues. An incident reported in the Forbes magazine raises an alarm over patient privacy [42]. In the report, it mentioned that Target Corporation sent baby care coupons to a teen-age girl unbeknown to her parents. This incident impels analytics and developers to consider privacy in big data. They should be able to verify that their applications conform to privacy agreements and that

sensitive information is kept private regardless of changes in applications and/or privacy regulations.

Privacy of medical data is then an important factor which must be seriously considered. We cite in the next paragraph some of laws on the privacy protection worldwide.

Data protection laws

More than ever it is crucial that healthcare organizations manage and safeguard personal information and address their risks and legal responsibilities in relation to processing personal data, to address the growing thicket of applicable data protection legislation. Different countries have different policies and laws for data privacy. Data protection regulations and laws in some of the countries along with salient features are listed in Table 2 below.

Privacy preserving methods in big data

Few traditional methods for privacy preserving in big data are described in brief here. Although these techniques are used traditionally to ensure the patient's privacy [43–45], their demerits led to the advent of newer methods.

A. De-identification

De-identification is a traditional method to prohibit the disclosure of confidential information by rejecting any information that can identify the patient, either by the first method that requires the removal of specific identifiers of the patient or by the second statistical method where the patient verifies himself that enough identifiers are deleted. Nonetheless, an attacker can possibly get more external information assistance for de-identification in big data. As a result, de-identification is not sufficient for protecting big data privacy. It could be more feasible through developing efficient privacy-preserving algorithms to help mitigate the risk of re-identification. The concepts of k-anonymity [46–48], l-diversity [47, 49, 50] and t-closeness [46, 50] have been introduced to enhance this traditional technique.

- *k-anonymity* In this technique, the higher the value of k, the lower will be the probability of re-identification. However, it may lead to distortions of data and hence greater information loss due to k-anonymization. Furthermore, excessive anonymization can make the disclosed data less useful to the recipients because some of the analysis becomes impossible or may produce biased and erroneous results. In k-anonymization, if the quasi-identifiers containing data are used to link with other publicly available data to identify individuals, then the sensitive attribute (like disease) as one of the identifier will be revealed. Table 3 is a non-anonymized database consisting of the patient records of some fictitious hospital in Casablanca.

There are six attributes along with five records in this data. There are two regular techniques for accomplishing k-anonymity for some value of k.

The first one is Suppression: in this method, an asterisk '*' could supplant certain values of the attributes. All or some of the values of a column may be replaced by '*'. In the

Table 2 Data protection laws in some of the countries

Country	Law	Salient features
USA	HIPAA Act Patient Safety and Quality Improvement Act (PSQIA) HITECH Act	Requires the establishment of national standards for electronic health-care transactions. Gives the right to privacy to individuals from age 12 through 18 Signed disclosure from the affected before giving out any information on provided healthcare to anyone, including parents Patient Safety Work Product must not be disclosed [63] Individual violating the confidentiality provisions is subject to a civil penalty Protect security and privacy of electronic health information
EU	Data Protection Directive	Protect people's fundamental rights and freedoms and in particular their right to privacy with respect to the processing of personal data [64]
Canada	Personal Information Protection and Electronic Documents Act ('PIPEDA')	Individual is given the right to know the reasons for collection or use of personal information, so that organizations are required to protect this information in a reasonable and secure way [65]
UK	Data Protection Act (DPA)	Provides a way for individuals to control information about themselves Personal data shall not be transferred to a country or territory outside the European Economic Area unless that country or territory ensures an adequate level of protection for the rights and freedoms of data subjects
Morocco	The 09-08 act, dated on 18 February 2009	Protects the one's privacy through the establishment of the CNDP authority by limiting the use of personal and sensitive data using the data controllers in any data processing operation [66]
Russia	Russian Federal Law on Personal Data	Requires data operators to take "all the necessary organizational and technical measures required for protecting personal data against unlawful or accidental access"
India	IT Act and IT (Amendment) Act	Implement reasonable security practices for sensitive personal data or information. Provides for compensation to person affected by wrongful loss or wrongful gain. Provides for imprisonment and/or fine for a person who causes wrongful loss or wrongful gain by disclosing personal information of another person while providing services under the terms of lawful contract
Brazil	Constitution	The intimacy, private life, honor and image of the people are inviolable, with assured right to indigenization by material or moral damage resulting from its violation
Angola	Data Protection Law (Law no. 22/11 of 17 June)	With respect to sensitive data processing, collection and processing is only allowed where there is a legal provision allowing such processing and prior authorization from the APD is obtained

anonymized Table 4, replaced each of the values in the 'Name' attribute and all the values in the 'Religion' attribute by a '*'.

The second method is Generalization: In this method, individual values of attributes are replaced with a broader category. For instance, The Birth field has been generalized to the year (e.g. the value '21/11/1972' of the attribute 'Birth' may be supplanted by the year '1972'). The ZIP Code field has been also generalized to indicate the wider area (Casablanca).

Table 4 has 2-anonymity with respect to the attributes 'Birth', 'Sex' and 'ZIP Code' since for any blend of these attributes found in any row of the table there are always no less than two rows with those exact attributes. Each "quasi-identifier" tuple occurs in at least k records for a dataset with k-anonymity. k-anonymous data can still be helpless against attacks like unsorted matching attack, temporal attack, and complementary release attack [50, 51]. On the bright side, the complexity of rendering relations of private records k-anonymous, while minimizing the amount of information that is not

Table 3 A non-anonymized database comprising of the patient records

Name	Birth	Sex	ZIP code	Religion	Disease
Yasmine	12/03/1962	Female	20502	Muslim	Heart-related
Khalid	21/11/1962	Male	20042	Muslim	Cancer
John	01/08/1964	Male	20056	Christian	Viral infection
Aicha	30/01/1962	Female	29004	Muslim	Diabetes mellitus
Abraham	15/09/1964	Male	20303	Jewish	Pneumonia

Table 4 2-anonymity with respect to the attributes 'Birth', 'Sex' and 'ZIP code'

Name	Birth	Sex	ZIP code	Religion	Disease
*	1962	Female	20000	*	Heart-related
*	1962	Male	20000	*	Cancer
*	1964	Male	20000	*	Viral infection
*	1962	Female	20000	*	Diabetes mellitus
*	1964	Male	20000	*	Pneumonia

released and simultaneously ensure the anonymity of individuals up to a group of size k, and withhold a minimum amount of information to achieve this privacy level and this optimization problem is NP-hard [52].

Various measures have been proposed to quantify information loss caused by anonymization, but they do not reflect the actual usefulness of data [53, 54]. Therefore, we move towards L-diversity strategy of data anonymization.

- *L-diversity* It is a form of group based anonymization that is utilized to safeguard privacy in data sets by diminishing the granularity of data representation. This model (Distinct, Entropy, Recursive) [46, 47, 51] is an extension of the k-anonymity which utilizes methods including generalization and suppression to reduce the granularity of data representation in a way that any given record maps onto at least k different records in the data. The l-diversity model handles a few of the weaknesses in the k-anonymity model in which protected identities to the level of k-individuals is not equal to protecting the corresponding sensitive values that were generalized or suppressed. The problem with this method is that it depends upon the range of sensitive attribute. If want to make data L-diverse though sensitive attribute has not as much as different values, fictitious data to be inserted. This fictitious data will improve the security but may result in problems amid analysis. As a result, L-diversity method is also a subject to skewness and similarity attack [51] and thus can't prevent attribute disclosure.
- *T-closeness* Is a further improvement of l-diversity group based anonymization. The t-closeness model (equal/hierarchical distance) [46, 50] extends the l-diversity model by treating the values of an attribute distinctly, taking into account the distribution of data values for that attribute. The main advantage of this technique is that it intercepts attribute disclosure, and its problem is that as size and variety of data increase, the odds of re-identification increase too.

B. HybrEx

Hybrid execution model [55] is a model for confidentiality and privacy in cloud computing. It utilizes public clouds only for an organization's non-sensitive data and computation classified as public, i.e., when the organization declares that there is no privacy and confidentiality risk in exporting the data and performing computation on it using public clouds, whereas for an organization's sensitive, private data and computation, the model executes their private cloud. Moreover, when an application requires access to both the private and public data, the application itself also gets partitioned and runs in both the private and public clouds. It considers data sensitivity before a job's execution and provides integration with safety.

The four categories in which HybrEx MapReduce enables new kinds of applications that utilize both public and private clouds are as shown in Fig. 2:

1. *Map hybrid (1a)* The map phase is executed in both the public and the private clouds while the reduce phase is executed in only one of the clouds.
2. *Vertical partitioning (1b)* Map and reduce tasks are executed in the public cloud using public data as the input, shuffle intermediate data amongst them, and store the result in the public cloud. The same work is done in the private cloud with private data. The jobs are processed in isolation.
3. *Horizontal partitioning (1c)* The map phase is executed only in public clouds, while the reduce phase is executed in a private cloud.
4. *Hybrid (1d)* The map phase and the reduce phase are executed on both public and private clouds. Data transmission among the clouds is also possible.

The problem with HybridEx is that it does not deal with the key that is generated at public and private clouds in the map phase and that it deals only with cloud as an adversary [55].

C. Identity based anonymization

It is a type of information sanitization whose intent is privacy protection. It is the process of either encrypting or removing personally identifiable information from data sets, so that the people whom the data describe remain anonymous. The main difficulty with this technique involves combining anonymization, privacy protection, and big data techniques [56] to analyze usage data while protecting the identities.

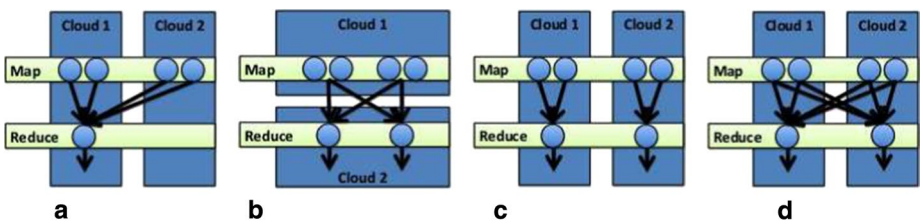

Fig. 2 The four Execution categories for HybrEx MapReduce [62]. **a** Map hybrid. **b** Horizontal partitioning. **c** Vertical partitioning. **d** Hybrid

Intel Human Factors Engineering team needed to protect Intel employees' privacy using web page access logs and big data tools to enhance convenience of Intel's heavily used internal web portal. They were required to remove personally identifying information (PII) from the portal's usage log repository but in a way that did not influence the utilization of big data tools to do analysis or the ability to re-identify a log entry in order to investigate unusual behavior.

To meet the significant benefits of Cloud storage [57], Intel created an open architecture for anonymization [56] that allowed a variety of tools to be utilized for both de-identifying and re-identifying web log records. In the implementing architecture process, enterprise data has properties different from the standard examples in anonymization literature [58]. Intel also found that in spite of masking obvious Personal Identification Information like usernames and IP addresses, the anonymized data was defenseless against correlation attacks. After exploring the tradeoffs of correcting these vulnerabilities, they found that User Agent information strongly correlates to individual users. This is a case study of anonymization implementation in an enterprise, describing requirements, implementation, and experiences encountered when utilizing anonymization to protect privacy in enterprise data analyzed using big data techniques. This investigation of the quality of anonymization used k-anonymity based metrics. Intel used Hadoop to analyze the anonymized data and acquire valuable results for the Human Factors analysts [59, 60]. At the same time, it learned that anonymization needs to be more than simply masking or generalizing certain fields—anonymized datasets need to be carefully analyzed to determine whether they are vulnerable to attack.

Summary on recent approaches used in big data privacy

In this paper, we have investigated the security and privacy challenges in big data, by discussing some existing approaches and techniques for achieving security and privacy in which healthcare organizations are likely to be highly beneficial. In this section, we focused on citing some approaches and techniques presented in different papers with emphasis on their focus and limitations (Table 5). Paper [61] for example, proposed privacy preserving data mining techniques in Hadoop. Paper [67] introduced also an efficient and privacy-preserving cosine similarity computing protocol and paper [68] discussed how an existing approach "differential privacy" is suitable for big data. Moreover, paper [69] suggested a scalable approach to anonymize large-scale data sets. Paper [70] proposed various privacy issues dealing with big data applications, while paper [71] proposed an anonymization algorithm to speed up anonymization of big data streams. In addition, paper [72] suggested a novel framework to achieve privacy-preserving machine learning and paper [73] proposed methodology provides data confidentiality and secure data sharing. All these techniques and approaches have shown some limitations.

These increased complexity and limits make the new models more difficult to interpret and their reliability less easy to assess, compared to previous models.

Conclusion

Whereas the potential opportunities offered for big data in the healthcare arena are unlimited (e.g. drive health research, knowledge discovery, clinical care, and personal health management), there are several obstacles that impede its true potential, including

Table 5 Summary on recent approaches used in big data privacy

Paper	Focus	Limitations
[56]	Discusses experiences and issues encountered when successfully combined anonymization, privacy protection, and Big data techniques to analyze usage data while protecting the identities of users	It still uses K-anonymity technique which is vulnerable to correlation attack
[61]	Proposed the privacy preserving data mining techniques in Hadoop, i.e. solve privacy violation without utility degradation	Its execution time is affected by noise size
[67]	Introduced an efficient and privacy-preserving cosine similarity computing protocol	Need significant research efforts for addressing unique privacy issues in some specific big data analytics
[68]	Discussed and suggested how an existing approach "differential privacy" is suitable for big data	This method depends totally on calculation of the amount of noise by the curator. So, if curator is compromised the whole system fails
[69]	Proposed a scalable two-phase top-down specialization (TDS) approach to anonymize large-scale data sets using the MapReduce framework on cloud	It uses anonymization technique which is vulnerable to correlation attack
[70]	Proposed various privacy issues dealing with big data applications	Customer segmentation and profiling can easily lead to discrimination based on age gender, ethnic background, health condition, social, background, and so on
[71]	Proposed an anonymization algorithm (FAST) to speed up anonymization of big data streams	Further research required to design and implement FAST in a distributed cloud-based framework in order to gain cloud computation power and achieve high scalability
[72]	The novel framework proposed into achieve privacy-preserving machine learning	The training data are distributed and each shared data portion of large volume, is not able to achieve distributed feature selection
[73]	Proposed methodology provides data confidentiality, secure data sharing without Re-encryption and access control for malicious insiders and forward and backward access control	Limiting the trust level in the cryptographic server

technical challenges, privacy and security issues and skilled talent. Big data security and privacy are considered huge obstacles for researchers in this field.

In this paper, we have briefly discussed some successful related work across the world. We have also presented privacy and security issues in each phase of big data lifecycle along with the advantages and flaws of existing technologies in the context of big healthcare data privacy and security.

We mainly reviewed the privacy preservation methods that have been used recently in healthcare and discussed how encryption and anonymization methods have been used for health care data protection as well as presented their limitations. Additionally, there are more various techniques include hiding a needle in a haystack [61], Attribute based encryption Access control, Homomorphic encryption, Storage path encryption and so on. However, the problem is always imposed.

In this context, as our future direction, perspectives consist in achieving effective solutions in privacy and security in the era of big healthcare data. As well, privacy methods need to be enhanced.

Also with the rapid development of IoT, the greater the quantity, the lower the quality. Consequently, quality of data should not be affected more by privacy preserving algorithms to get the appropriate result by researchers. And to go further, we will try

to solve the problem of reconciling security and privacy models by simulating diverse approaches to ultimately support decision making and planning strategies.

Abbreviations
NHS: National Health Service; UNCHC: UNC Health Care; PHI: Protected Health Information; ACMG: American College of Medical Genetics and Genomics; HCQI: Health Care Quality Indicators; OECD: Organisation for Economic Co-operation and Development; CCW: The Chronic Conditions Data Warehouse; MITM: man in the middle; TLS: transport layer security; SSL: secure sockets layer; VoIP: voice-over-IP; SHA: secure hash algorithm; RSA: Rivest Shamir and Adleman encryption algorithm; AES: advanced encryption standard; RC6: Rivest cipher 6; DES: data encryption standard; 3DES: triple data encryption algorithm; RC4: Rivest cipher 4; IDEA: international data encryption algorithm; RBAC: role-based access control; ABAC: attribute-based access control; EHR: electronic health record; CP-ABE: ciphertext-policy attribute-based encryption; DNS: domain name servers; IP: internet protocol; HTTP: hypertext transfer protocol; HIPAA: Health Insurance Portability and Accountability Act; PSQIA: Patient Safety and Quality Improvement Act; HITECH: Health Information Technology for Economic and Clinical Health; PIPEDA: Personal Information Protection and Electronic Documents Act; DPA: Data Protection Act; IT Act: Information Technology Act; CNDP: National Commission for Public Debate; APD: Agência de Proteção de Dados; PII: personally identifying information; TDS: two-phase top-down specialization; IoT: internet of things.

Authors' contributions
HK carried out the big data security studies in healthcare, participated in many conferences, the last one is The 8th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN 2017) in Lund, Sweden. She drafted also several manuscripts like "Big data security and privacy in healthcare: A Review" that was published in Procedia Computer Science journal. KA carried out the cloud computing security studies, participated in many conferences and drafted several manuscripts. Among these manuscripts, we find: "Assessing Cost and Response Time of a Web Application Hosted in a Cloud Environment" paper that was published by Springer in 2016. ABH carried out the cloud computing security studies, participated in many conferences and drafted multiple manuscripts as "Homomorphic encryption applied to secure storage and treatments of data in cloud" that was published in International Journal of Cloud Computing (IJCC), in 2016. All authors read and approved the final manuscript.

Authors' information
Not applicable.

Acknowledgements
The author forwards his heartfelt gratitude to two anonymous reviewers for their careful reading of the manuscript and their helpful comments that improve the presentation of this work.

Competing interests
The authors declare that they have no competing interests.

Availability of data and materials
Not applicable.

Consent for publication
Authors prove consent of publication for this research.

Ethics approval and consent to participate
Not applicable.

Funding
Not applicable (No payment is due on publication of this article. The article processing charge has been waived by Springer Open).

Publisher's Note

Received: 22 August 2017 Accepted: 20 December 2017
Published online: 09 January 2018

References

1. Burghard C. Big data and analytics key to accountable care success. Framingham: IDC Health Insights; 2012.
2. Fernandes L, O'Connor M, Weaver V. Big data, bigger outcomes. J AHIMA. 2012;83:38–42.
3. David Houlding, MSc, CISSP. Health Information at Risk: Successful Strategies for Healthcare Security and Privacy. Healthcare IT Program Of ce Intel Corporation, white paper. 2011.
4. South Tyneside NHS Foundation Trust. Harnessing analytics for strategic planning, operational decision making and end-to-end improvements in patient care. IBM Smarter Planet brief. 2013.

5. UNC Health Care relies on analytics to better manage medical data and improve patient care. IBM Press release. 2013.
6. Indiana Health Information Exchange. http://www.ihie.org/. Accessed 24 Mar 2016.
7. Transforming healthcare through big data, strategies for leveraging big data in the healthcare industry. Institute for Health. 2013.
8. Artemis. http://hir.uoit.ca/cms/?q=node/24. Accessed 21 May 2016.
9. Groves P, Kayyali B, Knott D, Kuiken SV. The big data revolution in healthcare, accelerating value and innovation. 2013.
10. WHO. Mobile phones help people with diabetes to manage fasting and feasting during Ramadan. Features. 2014.
11. Sophia Genetics. «Product & Technology Overview» 2014.
12. Sophia Genetics. http://www.sophiagenetics.com/news/media-mix/details/news/african-hospitals-adopt-sophia-artificial-intelligence-to-trigger-continent-wide-healthcare-leapfrogging-movement.html. Accessed 24 Mar 2017.
13. CynergisTek, Redspin. «BREACH REPORT 2016: Protected Health Information (PHI)» 2017.
14. Podesta J, et al. Big data: seizing opportunities, preserving values. Executive Office of the President. 2014;1:2013.
15. House W. Big data and privacy: a technological perspective. Washington: Executive Office of the President, President's Council of Advisors on Science and Technology; 2014.
16. House W. FACT SHEET: big data and privacy working group review. 2014.
17. OECD. Data-driven healthcare innovation, management and policy, DELSA/HEA(2013)13. Paris: OECD; 2013.
18. Kim S-H, Kim N-U, Chung T-M. Attribute relationship evaluation methodology for big data security. In: 2013 international conference on IT convergence and security (ICITCS), IEEE. p. 1–4. https://doi.org/10.1109/icitcs.2013.6717808.
19. "Data-driven healthcare organizations use big data analytics for big gains" IBM white paper February. 2013.
20. Yazan A, Yong W, Raj Kumar N. Big data life cycle: threats and security model. In: 21st Americas conference on information systems. 2015.
21. Xu L, Jiang C, Wang J, Yuan J, Ren Y. Information security in big data: privacy and data mining. J Rapid Open Access Publ. 2014;2:1149–76. https://doi.org/10.1109/ACCESS.2014.2362522.
22. General Dynamics Health Solutions white paper UK. "Securing Big Health Data"©2015. http://gdhealth.com/globalassets/health-solutions/documents/brochures/securing-big-health-data_-white-paper_UK.pdf.
23. Zhang R, Liu L. Security models and requirements for healthcare application clouds. In: IEEE 3rd international conference on cloud computing. 2010.
24. Shafer J, Rixner S, Cox AL. The hadoop distributed filesystem: balancing portability and performance. In: Proceedings of 2010 IEEE international symposium on performance analysis of systems & software (ISPASS), March 2010, White Plain, NY. p. 122–33.
25. Yang C, Lin W, Liu M. A novel triple encryption scheme for hadoop-based cloud data security. In: Emerging intelligent data and web technologies (EIDWT), 2013 fourth international conference on. 2013. p. 437–42.
26. Federal Information Processing Standards Publication 197. Specification for the advanced encryption standards (AES). 2001.
27. Somu N, Gangaa A, Sriram VS. Authentication service in hadoop using one time pad. Indian J Sci Technol. 2014;7:56–62.
28. Fluhrer S, Mantin I, Shamir A. Weakness in the key scheduling algorithm of RC4. In: 8th annual international workshop on selected areas in cryptography, London: Springer-Verlag. 2001.
29. Sweeney L. Achieving k-anonymity privacy protection using generalization and suppression. Int J Uncertain Fuzziness Knowl Based Syst. 2002;10:571–88.
30. Samrati P. Protecting respondents identities in microdata release. IEEE Trans Knowl Data Eng. 2001;13:1010–27.
31. Truta TM, Vinay B. Privacy protection: p-sensitive k-anonymity property. In: Proceedings of 22nd international conference on data engineering workshops. 2006. p. 94.
32. Spruill N. The confidentiality and analytic usefulness of masked business microdata. In: Proceedings on survey research methods. 1983. p. 602–607.
33. Chawala S, Dwork C, Sheny FM, Smith A, Wee H. Towards privacy in public databases. In: Proceedings on second theory of cryptography conference. 2005.
34. Science Applications International Corporation (SAIC). Role-based access control (RBAC) Role Engineering Process Version 3.0. 2004.
35. Mohan A, Blough DM. An attribute-based authorization policy framework with dynamic conflict resolution. In: Proceedings of the 9th symposium on identity and trust on the internet. 2010.
36. Hagner M. Security infrastructure and national patent summary. In: Tromso telemedicine and eHealth conference. 2007.
37. Zhou H, Wen Q. Data security accessing for HDFS based on attribute-group in cloud computing. In: International conference on logistics engineering, management and computer science (LEMCS 2014). 2014.
38. Linden H, Kalra D, Hasman A, Talmon J. Inter-organization future proof HER systems—a review of the security and privacy related issues. Int J Med Inform. 2009;78:141–60.
39. Marchal S, Xiuyan J, State R, Engel T. "A big data architecture for large scale security monitoring", Big Data (BigData Congress), Anchorage, AK. 2014. p. 56–63.
40. Duygu ST, Ramazan T, Seref S. A survey on security and privacy issues in big data. In: The 10th international conference for internet technology and secured transactions (ICITST-2015).
41. Liu L, Lin J. Some special issues of network security monitoring on big data environments. Dependable, Autonomic and Secure Computing (DASC), Chengdu. 2013. p. 10–5.
42. Hill K. How target figured out a teen girl was pregnant before her father did. Forbes, Inc. 2012.
43. Big Data security and privacy issues in healthcare—Harsh KupwadePatil, Ravi Seshadri. 2014.
44. Sectorial healthcare strategy 2012–2016-Moroccan healthcare ministry.
45. Patil P, Raul R, Shroff R, Maurya M. Big data in healthcare. 2014.
46. Li N, et al. t-Closeness: privacy beyond k-anonymity and L-diversity. In: Data engineering (ICDE) IEEE 23rd international conference. 2007.

47. Ton A, Saravanan M. Ericsson research. .http://www.ericsson.com/research-blog/data-knowledge/big-data-privacy-preservation/2015.
48. Samarati P. Protecting respondent's privacy in microdata release. IEEE Trans Knowl Data Eng. 2001;13(6):1010–27.
49. Machanavajjhala A, Gehrke J, Kifer D, Venkitasubramaniam M. L-diversity: privacy beyond k-anonymity. In: Proc. 22nd international conference data engineering (ICDE). 2006. p. 24.
50. Samarati P, Sweeney L. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical Report SRI-CSL-98-04, SRI Computer Science Laboratory. 1998.
51. Sweeney L. K-anonymity: a model for protecting privacy. Int J Uncertain Fuzziness. 2002;10(5):557–70.
52. Meyerson A, Williams R. On the complexity of optimal k-anonymity. In: Proc. of the ACM Symp. on principles of database systems. 2004.
53. Iyenger V. Transforming data to satisfy privacy constraints. In: Proceedings of the ACM SIGKDD. 2002;279–88.
54. LeFevre K, Ramakrishnan R, DeWitt DJ. Mondrian multidimensional k-anonymity. In: Proceedings of the ICDE. 2006. p. 25.
55. Priyank J, Manasi G, Nilay K. Big data privacy: a technological perspective and review. J Big Data. 2016;3:25.
56. Sedayao J, Bhardwaj R. Making big data, privacy, and anonymization work together in the enterprise: experiences and issues. In: Big data congress. 2014.
57. Yong Yu, et al. Cloud data integrity checking with an identity-based auditing mechanism from RSA. Future Gen Comput Syst. 2016;62:85–91.
58. Oracle big data for the enterprise. 2012. http://www.oracle.com/ca-en/technologies/biq-doto.
59. Hadoop Tutorials. 2012. https://developer.yahoo.com/hadoop/tutorial.
60. Fair scheduler guide. 2013. http://hadoop.apache.org/docs/r0.20.2/fair_scheduler.html.
61. Jung K, Park S, Park S. Hiding a needle in a haystack: privacy preserving Apriori algorithm in MapReduce framework PSBD'14, Shanghai. 2014. p. 11–7.
62. Ko SY, Jeon K, Morales R. The HybrEx model for confidentiality and privacy in cloud computing. In: 3rd USENIX workshop on hot topics in cloud computing, HotCloud'11, Portland. 2011.
63. Data Protection Laws of the World. 2017 DLA Piper. http://www.dlapiperdataprotection.com.
64. Privacy and Big Data—Terence Craig & Mary E. Ludloff.
65. Challenges of privacy protection in big data analytics—Meiko Jensen-2013 IEEE international congress on big data. 2013.
66. Data protection overview (Morocco)—Florence Chafiol-Chaumont and Anne-Laure Falkman. 2013.
67. Lu R, Zhu H, Liu X, Liu JK, Shao J. Toward efficient and privacy-preserving computing in big data era. IEEE Netw. 2014;28:46–50.
68. Microsoft differential privacy for everyone. 2015. http://download.microsoft.com/…/Differential_Privacy_for_Everyone.pdf.
69. Zhang X, Yang T, Liu C, Chen J. A scalable two-phase top-down specialization approach for data anonymization using systems, in MapReduce on cloud. IEEE Trans Parallel Distrib. 2014;25(2):363–73.
70. Mehmood A, Natgunanathan I, Xiang Y, Hua G, Guo S. Protection of big data privacy. In: IEEE translations and content mining are permitted for academic research. 2016.
71. Mohammadian E, Noferesti M, Jalili R. FAST: fast anonymization of big data streams. In: ACM proceedings of the 2014 international conference on big data science and computing, article 1. 2014.
72. Xu K, Yue H, Guo Y, Fang Y. Privacy-preserving machine learning algorithms for big data systems. In: IEEE 35th international conference on distributed systems. 2015.
73. Wei L, Zhu H, Cao Z, Dong X, Jia W, Chen Y, Vasilakos AV. Security and privacy for storage and computation in cloud computing. Inf Sci. 2014;258:371–86.