American Journal of Engineering Research (AJER) ***2017***
American Journal of Engineering Research (AJER)
e-ISSN: 2320-0847 p-ISSN : 2320-0936
Volume-6, Issue-7, pp-173-180
www.ajer.org
Research Paper Open Access

A Hybrid Privacy Preservation Framework for Healthcare Data Publishing

Kingsford Kissi Mireku[1], Zhang FengLi[2], Kittur Philemon Kibiwott[3]
[1, 2, 3](*School of Information and Software Engineering, University of Electronic, Science and Technology of China, China)*
Corresponding Author: Kingsford Kissi Mireku

Abstract: *While cloud computing and the emergence of big data phenomena presents significant opportunity for healthcare, they also elicit privacy concerns. These concerns indicate the need to ensure privacy protection when sharing or publishing the data. Currently, there are various approaches for addressing the problem of privacy protection, but these approaches have inherent weaknesses such as limitation in scope and reliance on privacy policies and guidelines that ultimately lead to insufficient protection or excessive distortion of data. In this paper, we propose a hybrid system that combines Map Reduce with k-Optimize algorithm to provide robust anonymity of health care data during publishing. The solution splits the data into distinct groups as specified by the k–Optimize algorithm and stored in anonymized data store. The major contribution of the paper is to demonstrate the feasibility of the hybrid approach and show its ability to maintain the consistency of the data consistent with best practices in big data management by preserving the privacy of patients' health data using k–Optimize algorithm and Map Reduce*

Keywords: *cloud computing; data sharing; healthcare data, k-Optimize, Map Reduce, privacy protection。*

Date of Submission: 08-07-2017 Date of acceptance: 15-07-2017

I. INTRODUCTION

With the recent development in information technology (IT) and the collection of large volumes of electronic information by data owners, data sharing has increased. This is especially important as more entities and organizations show greater willingness to collaborate driven by regulatory requirements and the mutual benefits associated with knowledge sharing and information-based decision-making [1]. Indeed, the evolution of IT has created an opportunity for information sharing and knowledge management in various fields [1].

In healthcare, the use of information and communication technology (ICT) has increased due to the potential improvements in healthcare effectiveness and efficiency. Cloud computing and big data have emerged as important disruptive technologies that influence health research communities. In particular, cloud computing provides robust computational power and allows massive storage capacity, which enables the deployment of applications without significant investment in the underlying infrastructure. When combined with cloud computing, big data results into large and complex datasets [2]. The large datasets raises a challenge for the conventional data processing tools in terms of handling the analysis of massive volumes of data derived from multiple sources. This has led to the incorporation of novel tools such as MapReduce to process large datasets [2].

Despite the increasing usage of collaborative IT and information sharing, privacy concerns pose a major challenge in the widespread use of these practices in healthcare [1]. In many cases, healthcare metadata contains privacy-sensitive information about individuals. Publishing this data could violate the principles of individual privacy. Currently, efforts to protect individual privacy focus on using guidelines and policies that restrict the type of published data as well as agreements on aspects such as storage of sensitive information. However, this approach has a number of limitations [3]. Notably, relying on policies and guidelines can lead to excessive distortion of data or require unfeasible level of trust in various scenarios of data sharing [3].

Various techniques and methods have been proposed to preserve privacy during data publishing without losing its utility. These methods include anonymization methods and encryption methods [4]. Anonymization techniques tend to use generalization and suppression (A-GS) algorithm, an approach that relies on dynamic programming as an approach to construct anonymized trajectories. A-GS entail recoding or replacing values with less other consistent values or using quasi identifiers, respectfully [4]. On the other hand,

encryption ensures the privacy of sensitive data by converting it from one form into another using cryptographic algorithm.

An overview of big data anonymization lifecycle on cloud is shown in Fig. 1 as describe by Zhang et al. [5]. They describe the lifecycle of the big data anonymization on cloud as follows. Phase (a) denotes data collection phase, where data owners submit their data into the data holders' applications. This describes data owners as any individual whose data is submitted and authorized its access. The main concern of the data owner is the privacy preservation of his data that he has authorized for analysis or sharing.

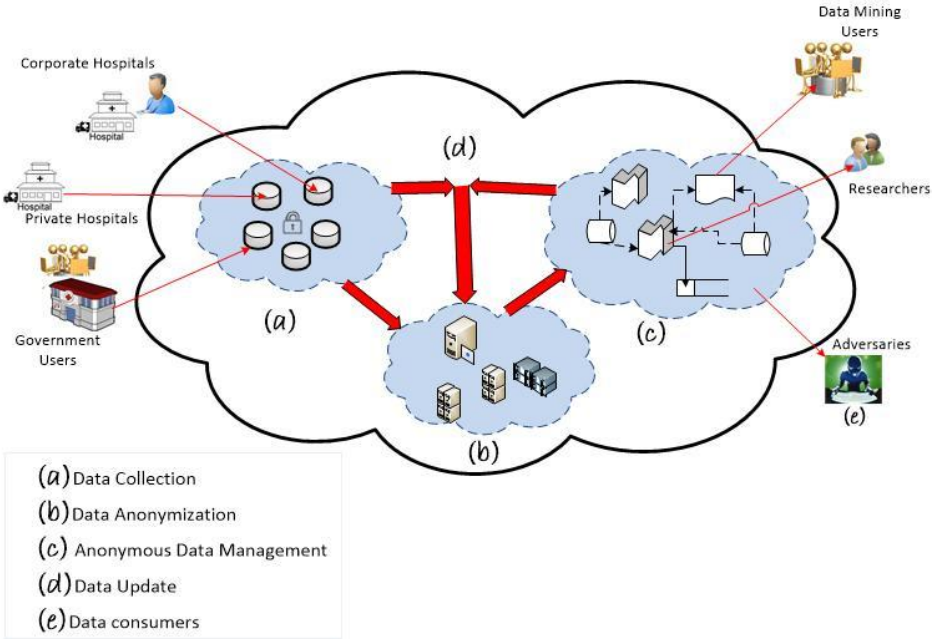

Figure 1. Overview of data anonymization lifecycle on cloud

The paper stated that phase (b) which is represented as data anonymization. With the assumption that anonymization services are reliable, the anonymization services will identify privacy preservation as its constraint. Privacy preserving requirements will be specified as parameters to anonymization services. In this phase our proposed system framework combines with the phase (c) which mainly manages anonymous data sets to ensure privacy preservation. We use the k–Optimize algorithm and MapReduce to preserve privacy of the data from the data collectors. After anonymization, the anonymous data sets are released to cloud. In general, there will be multiple anonymous data sets because of the different data usage or data recipients. Phase (d) is responsible for updating anonymous data sets when new data records are added. To anonymize the newly added data, the anonymized data sets should be taken into account and the whole anonymized data sets are offered to researchers.

Privacy-preserving data publishing uses methods and tools both for publishing and preserving data privacy and base on the lifecycle of data anonymization on cloud in figure 1 above, there are three parties involved in the privacy problem in data mining or researching:

a. Information Owner - wants to discover knowledge from the data without compromising the confidentiality of the data;
b. Information Providers (record owners) - individuals who provide their personal information to the data owner and want their privacy to be protected;
c. Data User/Miner/Recipient - has access to the data released by the data owner and can conduct data mining on the published. Data miner is considered as potential privacy intruder.

The content of each section of the paper continues as follows: section II discusses the related work while the preliminary and problem analysis is presented in section III by deliberating on the Privacy Preserving Data Publishing (PPDP), Data Anonymization Techniques and the problem statement. Section IV contains a detailed presentation on the proposed framework, while in section V the privacy preservation in the data publishing phase is discussed. We finally conclude our work in section VI.

II. RELATED WORK

Privacy protection is crucial for effective data or information sharing as it aims to protect sensitive data before processing or for information republication. Consequently, the PPDP concept continues to elicit scholarly attention with various frameworks for privacy protection developed in the recent past. In a previous study [6], proposed a novel framework for extracting, clustering, and de-identifying and anonymizing text-based patient identifiers by integrating various approaches established for data privacy in health informatics. The most notable contribution of the proposed model is the use of meta-learning approach to extract privacy-sensitive health information from documents and using recursive partitioning to cluster patient documents based on medical concepts [6].

In a pioneering study, [7] introduced a formal privacy protection model called k-anonymity, which identifies the quasi identifier. The solution focuses on person-specific data. In another pioneering study, [8] developed a framework for privacy protection using local suppression anonymization approach. The framework seeks to amend the limitations of k-anonymity methods and its extensions in the context of trajectory data. The study findings show that the proposed model has the ability to improve the quality of anonymized data [8].

In [9], the authors introduced a framework for privacy preserving data sharing using a constraint-based mining technique to generate sample datasets for sharing. The proposed dataset reconstruction solution allows data owners to control the mine-able knowledge from the dataset. The main contribution of the dataset reconstruction method is that it demonstrates the integration of inverse itemset lattice mining as a solution to the problem of privacy-preserving data sharing [9].

In recognizing the need to balance privacy and utility, [10] developed a framework for data sharing in the Internet considering both data utility and user privacy. The Privacy-Sensitive Sharing framework or PS2 provides a robust, transparent, consistent, and replicable method for evaluating risks and benefits as opposed to relying on opaque and subjective decision metrics. The framework envisions that transparency issues surrounding information sharing aggravates privacy problems. For this reason, the researchers developed PS2 to improve transparency in data sharing [10].

With the use of numeric QI [11], the paper describes the record linkage and attribute linkage attacks in privacy preservation era. The paper further applied the generalization of the numeric QI in the process of specifying range values and also implementing the record elimination in achieving the privacy preservation of the data.

Another work addressing privacy preservation in publishing health records is presented in SaC-FRAPP: a scalable and cost-effective framework for privacy preservation over big data on cloud. [5] The key idea of the framework proposed by them is that it leverages cloud-based MapReduce to conduct data anonymization and manage anonymous data sets, before releasing data to others. The framework provides a holistic conceptual foundation for privacy preservation over big data.

A properly designed computerized third-party platform proposed by Hye-Chung Kum, and Ashok Krishnamurthy, [12] using SDLink, can precisely control the information disclosed at the micro level and allows frequent human interaction during the linkage process, is an effective human–machine hybrid system that can accurately and safely integrate Big Data for biomedical research. This technique record linkage to integrate uncoordinated databases is critical in biomedical research using Big Data. Balancing privacy protection against the need for high quality record linkage requires a human–machine hybrid system to safely manage uncertainty in the ever changing streams of chaotic Big Data. The problem here is that most linkage is done by staff which makes it easy to leak information to adversaries.

Recently, Roberto J. Bayardo and Rakesh Agrawal [13] studied Data Privacy through Optimal k-Anonymization and stated that “A desirable feature of protecting privacy through k-anonymity is its preservation of data integrity”. Indeed, it may be interesting to consider combined approaches, such as k-anonymizing over only a subset of potentially identifying columns and randomly perturbing the others. The problem is a better understanding of when and how to apply various privacy-preserving methods? They proposed Optimal algorithms to be useful in this regard since they eliminate the possibility that a poor outcome is the result of a highly sub-optimal solution rather than an inherent limitation of the specific technique.

III. PRELIMINARY AND PROBLEM ANALYSIS

A. Privacy Preserving Data Publishing

Privacy protection practices envisage that published data would not increase the risk of an adversary gaining additional information about individuals if when the adversary has the background knowledge [14]. The PPDP concept provides tools and methods for preserving data privacy when publishing information. Generally, the PPDP process encompasses two phases: data collection phase and data publishing phase. The PPDP process envisions three fundamental roles including the data owners, the data publisher, and the data recipient [14]. Fig. 2 illustrates the relationship between the two phases and the three roles in the general PPDP process. The process starts with data collection phase in which the data publisher collects healthcare datasets from a data

owner. In the second phase, the data publisher sends processed datasets to the data recipient. Essentially, it is not possible to send raw datasets from the data owner directly to the data recipient; that is, the data publisher needs to process the dataset before sending it to the data recipient [14].

An extension of the approach in [6] envisages dividing the data publisher into two model categories: trusted and untrusted model. In the trusted model, the data publisher plays a reliable or trustworthy role, which means that their data is safe and without any significant risk. On the contrary, the data publisher can gain privacy from dataset in the untrusted model. Various cryptographic methods and statistical techniques can help in collecting records anonymously from their owners without revealing their identities in the untrusted model.

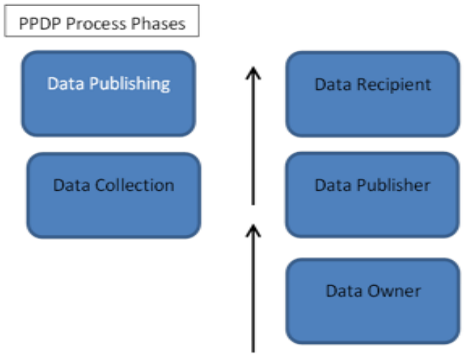

Figure 2. The relationship of PPDP process phases and roles [adopted from 5]

The classification of trust models based on the data publisher results into four distinct scenarios that influence PPDP. The first scenario affecting PPDP is the non-expert data publisher [3]. In this scenario, the data publisher does not require specific knowledge about research fields; they only need to make data meet the requirements of information privacy and data utility. The second scenario is where there is risk the data recipient could be an adversary, the requisite hypothesis in many practical applications [3]. The third scenario is that of a published data being not the result of the data mining, which indicates that the dataset that the data publisher provides is not for data mining purposes only [3]. The fourth scenario is that is truthfulness at record level. This scenario envisages the data publisher guaranteeing data authenticity during publishing irrespective of the processing methods deployed. This suggests that randomization techniques and perturbation methods cannot meet the privacy requirements in the scenario [3].

B. Data Anonymization Techniques

Anonymization is a PPDP approach that seeks to conceal sensitive data assuming that data analysis should not expose sensitive information [3]. Anonymization involves sanitizing data in order to protect private information and ensure privacy principles such as k-anonymity, I-diversity, as well as using generalization, suppression, and micro-aggregation. The process entails removing the explicit identifiers of data owners [3].

K-Anonymity

K-anonymity is an anonymization technique that assumes that each record in a dataset cannot be distinguished with another (k-1) records after anonymous operations when considering quasi identifiers [4]. K-anonymity can also be explain as a property of anonymized data [15]. i.e., data in which the identifiable information are hidden from data sets for the privacy preservation of the database. That is, k-anonymity guarantees the probability of representing an individual uniquely in datasets does not exceed 1/k. In the context of data anonymization, most data types, especially in relational databases, conform to the two-dimensional table. The attributes of such tables based on the need to preserve privacy fall into four distinct categories: identifier, quasi-identifiers, non-quasi attributes, and sensitive attributes [4]. The identifier attribute represents a unique individual. This attribute should be removed prior to data processing. The quasi-identifiers describe specific attribute sequences occurring in the tables that allow malicious attackers to link released datasets with other dataset to break privacy and gain sensitive information. The notion of data sanitization targets the quasi identifiers attribute. Since there is always uncertainty about the number of quasi identifiers, all PPDP approaches anticipate quasi-identifiers sequence beforehand [4]. The non-quasi attributes do not affect data processing. However, sensitive attributes tend to contain privacy-sensitive information such as disease or salary. This demonstrates why data processing approaches focus on quasi identifiers to reduce dataset correlations before publishing [4].

C. Problem Statement

Protecting the privacy of healthcare information remains an open problem, especially when sharing information or publishing the data. Indeed, big data and cloud computing technologies have enabled data owners to collect massive electronic information and provided opportunities for enhanced information retrieval and knowledge management [4]. While these developments have improved decision making, especially in the field of medical research and publishing, they also elicit significant privacy concerns. In particular, processing, sharing, and publishing healthcare data can lead to information misuse, disclosure of the identity of the data owner, and other related privacy violations. Therefore, the primary objective of privacy preservation is to ensure appropriate protection of sensitive data before analysis of publication. In order to achieve this, various techniques or algorithms have been proposed including the use of meta-learning approach [6], using k-anonymity and its extensions [7], using local suppression anonymization [8], and constrain-based mining techniques [9], among others. These techniques have their own limitations in terms of the quality of the anonymized data, the level of protection, and their practical feasibility. To address this problem, we propose a hybrid system that combines the utility of MapReduce with k-Optimize software to enable two levels of anonymity and make it unfeasible for adversaries to infringe data privacy in health records.

IV. PROPOSED FRAMEWORK

Healthcare information may contain privacy-sensitive data that may violate the privacy of individual patients or compromise the operations of health institutions. To protect the privacy of health care data during publishing, we propose a novel framework that uses an algorithm for splitting the data into distinct groups and subsequently storing the data into anonymized data store. Fig. 3 below illustrates the hybrid framework.

The privacy protection algorithm generates two tables: α-index and β-index an extension of the paper [16]. The α-index table consists of index mapping for all the healthcare information apart from the "identity-defining information" related to health records. On the other hand, β-index table contains an index of "identity defining information" related to health records, but as segregated by k-Optimize algorithm. The solution envisages creating a second level indexing by map reducing an Hbase table of anonymized data and α-index table. The Hbase table contains only the anonymized index records generated by k–Optimize algorithm. The solution envisages distribution of the data on HDFS cluster. Whenever a researcher queries the health care records, the search query is invoked on the α-index table, which essentially compares the original data in the data store. The Hbase table for indexing helps to fetch the mapping results from HDFS. This means that the final dataset would not reveal the original data as all the privacy-sensitive data would be separated before passing it to the HDFS. The solution allows publishing of precise datasets for research purposes while maintaining anonymity or privacy as defined by users [16].

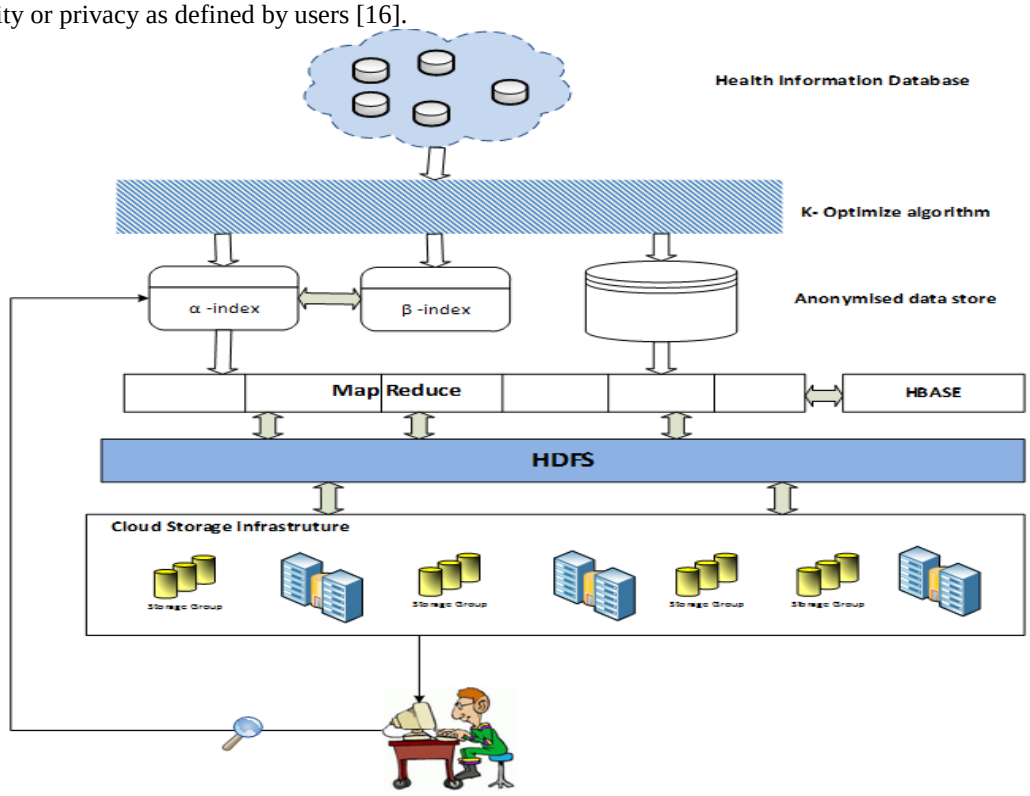

Figure 3. Proposed framework

A. Proposed K-Anonymity Approach

The hybrid system combines both MapReduce and k-optimize as a strategy to provide two levels of anonymity. This approach makes it practically impossible for adversaries to violate the privacy of sensitive heath records during publishing. The anonymization approach proposed for this study is to use k-anonymity in which the basic form of PPDP features a data publisher in a table of the following form [3]:

D(Explicit_ Identifier, Quasi-identifier, Sensitive_ Attributes, Non-Sensitive_ Attributes)

In the above case, the Explicit_ Identifier represents a set of attributes containing information that explicitly identifies a data owner. A Quasi_ Identifier or QID represents a set of attributes that potentially identify data owners. Sensitive_ Attributes comprise of privacy-sensitive information such as diseases and disability status, while Non-Sensitive_ Attributes contains attributes that do not fall under any of the other categories. In this anonymity model, it is assumed that the data publisher does not know the QID. Most existing models consider a single QID with attributes that can serve as quasi identifiers [3]. This means that the more attributes considered in QID, the more the protection provided by the k-anonymity. Similarly, this implies the need for additional distortion to attain k-anonymity because datasets must agree on various attributes. This specification means that if a data publisher wants to publish a table T (A, B, C, D, S), where S represents a sensitive attribute, and is aware that the data recipient has some form of access to the tables (T1 (A, B, X) and T2(C, D, Y). In which X and Y are attributes besides T. In order to prevent linking of the data records in T to the sensitive information on X and Y, the data publisher has the opportunity to specify k-anonymity on QID1 = {A, B) and QID2 = {C, D) for T. This approach suggests that each data record in T can be distinguished from a group of a minimum of k records regarding QID 2; that is, the two groups are not the same. This requirement is implied by k-anonymity on QID = {A, B, C, D), but with k-anonymity on both QID 1, and QID 2 [3].

This study envisages using the k-optimizing algorithm, which utilizes a sub-tree generalization scheme as well as record suppression [13]. The algorithm operates by pruning all non-optimal anonymous tables and modeling the search space based on a given set of enumeration tree.in this model, each node will represent a k-anonymous solution. The algorithm is based on the assumption that there exist ordered sets of attributes and thus examines the enumeration tree in a top-down approach [13]. The K-Optimize algorithm splits the data in health information database into appropriate groups and stores them in a data store, which ensures privacy during publishing.

B. Mapreduce Approach

Hadoop MapReduce can be define as an open-source software framework used for building a reliable, scalable, distributed applications in computing environment for processing large amount of data. [17]. By applying MapReduce, input data-set are split into chunks to be independent of each other. A map function task is later use to process the split chucks in parallel.

A framework of a MapReduce framework is made up of a single master JobTracker and one slave TaskTracker in which the function of the master JobTrackeris to be in charge of the scheduling of MapReduce jobs' module tasks on the slave nodes. This helps in observing the tasks on the slave nodes and at the same time identify failed tasks for re-execution. In the other hand, the slave TaskTracker is in charge of executing all tasks which has be assigned by the master JobTracker. [18].

With reference to Big data [19], MapReduce framework is being implement as <key, value> pairs, this enables the framework to show the input to the job as a set of <key, value> pairs and gives result as a set of <key, value> pairs.

The proposed solution targets the MapReduce paradigm, which comprises of series of slave node computers and master computer. While u uploads files into the MapReduce cloud, each file in this architecture is split into blocks called *InputSplits.* These files have fixed sizes $S_{InputSplit}$, a pre-configured system parameter. If S_{File} symbolizes the size of an uploaded file, the number of InputSplits c computes to $c = S_{File}/S_{InputSplit}$. The MapRedice also allows u to upload operations in complied Java classes besides data [20]. The classes symbolize the implementation of three functions [20]:

Overall procedure will follow four primary steps described in [21]:

1. Scan (INPUTSPLIT) ← [(k; v)]), this functions takes InputSplit as input and parses it to create a set of key-value pairs [(k; v)].
2. Map (k; v) ← [(k0; v0)], this function takes a single key-value pair as input (k; v) and generates intermediate key-value pair sets [(k0; v0)].
3. Reduce ([(k0; v0)]) ←FILE, this operation the intermediate key-value pairs as input [(k0; v0)] to create arbitrary output into file.

1. Preparation of the Map () input
2. Running the user-provided Map () code once for each K1 value, and generating the output as organized by key values K2
3. Shuffling the Map output to the Reduce Processors
4. Running the user-provided Reduce () code for each K2 key value generated in the Map step

The idea of combining k-anonymity and MapReduce algorithm is to ensure robust privacy while maintaining data consistency.

V. PRIVACY PRESERVATION IN THE DATA PUBLISHING PHASE

From the big data anonymization lifecycle on cloud in Fig. 1, the publishing of the analyzed data on cloud for sharing can be done using Anonymization techniques to maintain the privacy of the data which is processed. The Anonymization techniques are done at the data anonymization phase (b) by using the k–Optimize algorithm to generate two tables: α-index and β-index where the α-index table contains the index mapping for all the healthcare data while the β-index table consists of "identity defining information" related to health records. By creating a second level of anonymization, MapReduce applies the Hbase table to distribute the anonymized data and α-index table generated by k–Optimize algorithm on HDFS cluster. This protects the privacy of the data owners while publishing their data on big data platform.

A. Security Analysis

Attacking model - An adversary may be able to launch the following attacks;

1. Homogeneity attack - this attack affects k-anonymity algorithm in which the sensitive values within a set of k records are indistinguishable, the sensitive value can be identify. This attack indicates the leakage of information which is caused by the groups created by k-anonymity due to the lack of multiplicity in the sensitive attribute. [15]
2. Background knowledge attack - the attacker has background knowledge if there is a connotation between one or more Quasi-Identifiers (QI) attributes within the Sensitive attributes (SA) to reduce the set of possible values for the Sensitive attributes (SA) [15].

With these attacks, one can conclude that privacy preservation for health data publishing on cloud are unprotected since k-anonymous table may release sensitive information. The proposed hybrid framework disproof the above conclusion by demonstrating a two level anonymity by combining k-Optimized algorithm and MapReduce algorithm in ensuring a robust privacy while maintaining data consistency during the publishing of health data on cloud. This provides a resilient definition of privacy that considers diversity and background knowledge.

VI. CONCLUSION

Privacy Preserving Data Publishing offers enormous opportunities in healthcare because it fosters information sharing while protecting the privacy of sensitive information. The general objective in PPDP is to transform original health care data into an anonymous form to prevent potential privacy violations. In this paper, we reviewed the relationship involving PPDP phases and roles and the trust models in published health data. The research shows two models of data publishers: in the trusted model, the data publisher is a trustworthy entity and the data owners can contribute their personal information. In the untrusted model, the data publisher could also be an adversary who targets data owners and their sensitive information. We also reviewed existing literature describing frameworks privacy protection and identified their limitations.

This paper proposes a framework which combines k-anonymity and MapReduce algorithm to ensure robust privacy while maintaining data consistency. The solution allows publishing of precise datasets for research purposes while maintaining anonymity or privacy of users.

REFERENCES

[1] A. H. Rashid and N. B. M. Yasin. "Sharing healthcare information based on privacy preservation". Academic Journals, vol. 10. No. 5, pp. 184-195.
[2] V. N. Inukollu, S. Arsi, and S. R. Ravuri. "Security issues associated with big data in cloud computing". International Journal of Network Security and Its Applications, vol. 6, no. 3, 2014, 45-56.
[3] B. C. Fung, K. Wang, R. Chen, and P. S. Yu. "Privacy-preserving data publishing: a survey of recent developments". ACM Computing Surveys, vol. 42, no. 4, 2010.
[4] B. Selvaraj and S. Periyasamy. "A review of recent advances in privacy preservation in health care data publishing". International Journal of Pharma and Biosciences, vol. 7, no. 4, pp. 33-41.
[5] Xuyun Zhang, Chang Liu, Surya Nepal, Chi Yang, Wanchun Dou and Jinjun Chen "SaC-FRAPP: a scalable and cost-effective framework for privacy preservation over big data on cloud" Concurrency Computat.: Pract. Exper. 2013
[6] X. B. Li and J. Qin. "A framework for privacy-preserving medical document sharing". 34th International Conference on Information Systems, Milan, 2013, pp. 1-17.

[7] L. Sweeney. “K-Anonymity: a model for protecting privacy”. International Journal on Uncertainty Fuzziness and Knowledge-based Systems, vol. 10, no. 5, 2002, pp. 557-570.
[8] R. Chen, B. C. M. Fung, N. Mohammed, Desai, B. C. and K. Wang. “Privacy-Preserving trajectory data publishing by local suppression”. Data Mining for Information Security, Information Sciences, vol. 231, no. 10. 2013, pp. 83-97.
[9] X. Chen, M. Orlowska, and X. Li. “A new framework of privacy preserving data sharing”. In Proceedings of the IEEE 4th International Conference on Data Mining ICDM04 Workshop: Privacy and Security Aspects of Data Mining, Brighton, UK, 1-4 November, 2004, pp. 47-56.
[10] E. E. Kenneally and K. Claffy. “An internet data sharing framework for balancing privacy and utility”. 1st International Forum on the Application and Management of Personal Electronic Information, 2009, MIT.
[11] Mahesh, T. Meyyappan. “Anonymization Technique through Record Elimination to Preserve Privacy of Published Data”, Proceedings of the 2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering, February 21-22, R.
[12] Hye-Chung Kum, Ashok Krishnamurthy, AshwinMachanavajjhala, Michael K Reiter, Stanley Ahalt. “Privacy preserving interactive record linkage (PPIRL)”
[13] Roberto J. Bayardo IBM Almaden Rakesh Agrawal IBM Almaden. “Data Privacy Through Optimal k-Anonymization” Research Center rakesh_agrawal@ieee.org
[14] Y. Xu. T. Ma, M. Tang, and W. Tian. “A survey of privacy preserving data publishing using generalization and suppression”. Applied Mathematics and Information Sciences, vol. 8, no. 3, pp. 1103-1116.
[15] Balaji K. Bodkhe and Sanjay P. Sood. “Hadoop Used Medical Analytics and Privacy Preservation” International Journal of Control Theory and Applications, Vol 10, No. 30, 2017
[16] K. Mireku, Z. FengLi, M. Dennis, A. Khan and I. Khan. “Secured cloud database healthcare mining analysis”. IEEE, 2016, PP. 3937-3940.
[17] Apache Hive for Apache Hadoop, https://hive.apache.org and Apache Spark for Apache Hadoop, http://spark.apache.org/.
[18] Aditya B. Patel, Manashvi Birla, Ushma Nair, “Addressing Big Data Problem Using Hadoop and Map Reduce”, 2012,
[19] Big Data: A Review by Seref SAGIROGLU and Duygu SINANC Gazi University,IEEE 2014z
[20] E. O. Blass, R. Di Pietro, R. Molva, and M. Onen. “PRISM – Privacy-preserving search in MapReduce”. LNCS 7384, 2012, pp. 180-200.
[21] R. Sreedhar and Imamaheshwari, D. “Big data processing with privacy preserving MapReduce cloud”. International Journal of Innovative Research in Science, Engineering, and Technology, vol. 3, no. 1, 2014, pp. 343-350.

Kingsford Kissi Mireku. "A Hybrid Privacy Preservation Framework for Healthcare Data Publishing." American Journal of Engineering Research (AJER) 6.7 (2017): 173-180.