International Journal of Innovative Research in Science and Engineering
Vol. No.3, Issue 04, April 2017
www.ijirse.com
ISSN (O) 2454-9665
ISSN (P) 2454-0663

ANALYZING THE PRIVACY PRESERVING USING BIG DATA TECHNIQUES

M.C.S.Geetha[1], N.Selvakumar[2], W.Willfred Jose[3]

[1]*Research Scholar, Avinashilingam University (India)*

[2,3]*MCA Student, Kumaraguru College of Technology (India)*

ABSTRACT

Recently big data has become a hot research topic. The rising amounts of big data also increase the chance of violate the privacy of individuals. Since big data need high computational power and large storage, distributed systems are used. As multiple parties are concerned in these systems, the risk of privacy violation is improved. There have been a number of privacy-preserving methods developed for privacy protection at different stages (e.g., data generation, data storage, and data processing) of a big data life cycle. The goal of this paper is to provide a comprehensive overview of the privacy preservation methods in big data and present the challenge for existing mechanisms. In this paper, we illustrate the infrastructure of big data and the big data life cycle. Furthermore, we discuss the research challenges of the privacy preservation in big data. This paper also presents recent techniques of privacy preserving in big data like hiding a needle in a haystack, identity based anonymization, differential privacy and privacy-preserving big data publishing of big data streams and future research directions related to big data.

Keywords: Big data, privacy, data auditing, big data storage, big data processing.

I. INTRODUCTION

Due to recent scientific development, the quantity of data generated by social networking sites, Internet, sensor networks, healthcare applications is drastically increasing day by day. All the vast amount of data generated from dissimilar sources in multiple formats with very high speed is referred as big data. The data generation rate is growing so rapidly that it is becoming extremely difficult to handle it by means of traditional methods or systems [1]. Big data could be structured, semi-structured or unstructured which includes more challenges when performing data storage and processing tasks. Therefore we need new ways to analyze and store data in actual time. Big data has been captured and analyzed in a timely manner can be transformed into actionable imminent which can be of significant value. It can help businesses and organizations to progress the internal decision making power and can build new opportunities throughout data analysis. It can help to encourage the scientific research and nation by transforming traditional business models and scientific values [2].

Big data can be defined in different ways. Here we use the definition given by International Data Corporation (IDC) in [3]. The term big data is defined as ‘‘a new generation of technologies and architectures considered to economically extract value from very huge volumes of a widespread variety of data, by enabling high-velocity capture, discovery, and/or analysis’’. Based on this definition, the things of big data are reflected by 3 V’s,

ISSN (O) 2454-9665
ISSN (P) 2454-0663

which are, velocity, volume and variety. Volume refers to the amount of data generated. With the emergence of social networking sites, we have seen a remarkable increase in the size of the data. The rate at which new data are generated is often characterized as velocity. A common theme of big data is that the data are different, i.e., they may contain text, image, audio or video etc. This diversity of data is denoted by variety. Despite big data could be effectively utilized for us to well understand the world and invent in various aspects of human endeavors, the exploding amount of data has improved potential privacy breach. For example, Amazon and Google can learn our browsing habits and shopping preferences. Social networking sites such as Facebook collect all the information about our individual life and social relationships. Common video sharing websites such as YouTube mentions us videos based on our search history.

In 2006, AOL released 20 million queries for 650 users by eliminating the AOL id and IP address for research. However, it procured researchers only couple of days to re-identify the users. Users' privacy may be broken under the following circumstances [4]:

• Personal data when combined with external datasets may lead to the implication of new facts about the users. Those proofs may be secretive and not supposed to be discovered to others.

• Personal data is sometimes composed and used to add value to business. For example, individual's shopping habits may reveal a lot of personal information.

• The sensitive information are stored and processed in a location not secured properly and data leakage may occur during storage and processing phases.

In order to confirm big data privacy, several tools have been developed in recent years. These tools can be grouped based on the phases of big data life cycle, i.e., data generation, processing and storage. In data generation phase, for the safety of privacy, access constraint and misrepresenting data techniques are used. While access constraint techniques try to bounds the access to individuals' private data, forging data techniques alter the original data before they are unconfined to a non-trusted party.

The methods to privacy protection in data storing phase are primarily based on encryption techniques. Encryption techniques can be further separated into Identity based encryption (IBE), attribute based encryption (ABE) and storage path encryption. Hybrid clouds are used where sensitive information is stored in private cloud. The data processing stage includes knowledge extraction from the data and privacy preserving data publishing (PPDP). In PPDP, anonymization techniques such as suppression and generalization are used to safeguard the privacy of data. Guaranteeing the utility of the data while protecting the privacy is a excessive challenge in PPDP.

In the knowledge mining process, there exist several tools to extract useful data from complex and large-scale data. These tools can be further distributed into classification, clustering and association rule mining techniques. While classification and clustering divided the input data into dissimilar groups, association rule mining techniques discover the useful relationships and developments in the input data. Protecting privacy in big data is a fast developing research area. Even though some related papers have been available but only few of them are survey type of papers [2], [5]. While these papers introduced the basic concept of privacy protection in big data, they failed to cover some important aspects of this area.

For example, neither [2] nor [5] be responsible for detailed deliberation regarding big data privacy with detail to cloud computing. None of the papers deliberated future challenges in detail. In this paper, we will give a

ISSN (O) 2454-9665
ISSN (P) 2454-0663

complete overview of the privacy preservation of big data at each stage of big data life cycle. Moreover, we will discuss privacy issues related to big data when they are stored and processed on cloud, as cloud computing plays very vital role in the application of big data. Furthermore, we will discuss about potential research directions. The remainder of this paper is organized as follows. The life cycle of the big data related to privacy will be discussed in section II. Research challenges related to big data will be discussed in section III. Recent techniques of privacy preserving in big data will be discussed in sections IV. Finally future research directions are identified in section V.

II. LIFE CYCLE OF BIG DATA

Data can be generated from numerous distributed sources. The quantity of data generated by machines and humans has blowed up in the past few years. For example, everyday 2.5 quintillion bytes of information are generated on the web and 90 percent of the information in the world is created in the past few years. Facebook, a social networking site alone is creating 25TB of new data every day. The data generated is huge, diverse and complex. It is hard for traditional systems to handle them. The information generated are associated with a specific domain such as business, Internet, research, etc.

• Data storage: This stage refers to storing and dealing large-scale data sets. A information storage system comprises of two parts i.e., hardware organization and data management [6]. Hardware organization refers to using information and communications technology (ICT) properties for various responsibilities. Data organization refers to the set of software deployed on top of hardware organization to succeed and query large scale data sets. It should provide numerous interfaces to interrelate with and analyze stored information.

• Data processing: Data processing stage refers to the process of information collection, pre-processing, data transmission and takes out useful information. Data collection is required because information may be coming from diverse sources i.e., sites that contains images, text and videos. In data collection stage, data are attained from specific data production location using dedicated data collection technology. In data transmission stage, after collecting raw data from a exact data production location we need a high speed transmission tool to transmit data into a appropriate storage for several type of analytic applications.

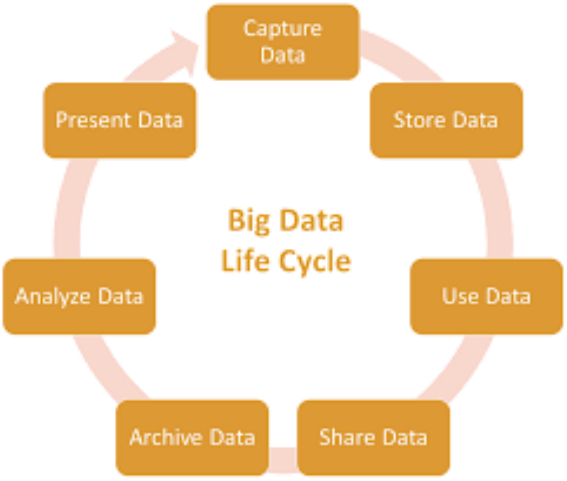

Fig 1. Illustration of big data life cycle.

ISSN (O) 2454-9665
ISSN (P) 2454-0663

Finally, the pre-processing stage aims at eliminating meaningless and redundant parts of the information so that more storage space could be protected. The unnecessary data and domain specific analytical techniques are used by many applications to develop meaningful information. Although diverse fields in data analytics require diverse data characteristics, few of these fields may control similar underlying technology to review, transform and model information to extract value from it. Emerging data analytics research can be categorized into the following six technical regions: structured data analytics, multimedia analytics, text analytics, web analytics, mobile analytics and network analytics [6].

III. RESEARCH CHALLENGES AND OPPORTUNITIES

A. Challenges of privacy

Big data is a new atmosphere for computer science, and privacy is one of the serious problem, which has to be suitably addressed before we can appreciate the pervasive applications of big data. There are many problems and challenges in terms of privacy study in big data. We review the major ones here for readers established on our current understanding. 1) Measurement of privacy. As privacy is a subjective concept, it various from person to person, from time to time even for the same person. It is hard to define it, and therefore, hard to measure. This problem is fundamental and challenging. It needs the effort not only from technical aspects, but more from social and psychological perspective.

2) Theoretical framework of privacy. We now have data clustering methods and the differential privacy framework for data privacy. However, we also see the limitations of various data clustering methods, and the needs to adapt the differential privacy in practice. Should we have new and better theoretical foundations for privacy study in big data era? we believe the answer is positive, and it takes time.

3) Scalability of privacy algorithms. We have some mechanisms and strategies in place to handle big databases, and the main strategy is divide and conquer. However, the scale of big data is far bigger than a database. Therefore, it is challenging to design scalable algorithms for privacy algorithms.

4) Heterogeneity of data source. The available privacy algorithms are almost all for homogeneous data sources, such as the records in databases. However, the data sources of forthcoming big data are heterogeneous with a high probability. It is challenging to deal with heterogenous data sources in an efficient way.

5) Efficiency of privacy algorithms. Given the volume of big data, efficiency becomes a very important element of privacy algorithms in the big data environment.

B. Opportunities

New privacy frameworks and mechanisms are highly expected in the near future. Based on our understanding, we believe the followings are the promising directions for our investment. 1) Quantum computing for unconditional privacy preserving. We are getting closer and closer to the practical usage of quantum computers. One good news is that quantum computing can offer fantastic functions in security and privacy preserving. The majority advantage of the current encryption methods is time complexity. However, it is not an unconditional method for privacy protection or security.

The recently proposed measurement based model of quantum computation [7] provide a promising diagram to achieve blind computing, meaning a client can delegate a computation to a quantum server, and the server can execute the task without gaining any knowledge about the input, output, and the client. Braz et al. [8]

ISSN (O) 2454-9665
ISSN (P) 2454-0663

implemented the conceptual framework of the model and demonstrated the feasibility of blind quantum computing. It is time to start our exploration in significant computing not only for privacy and security, but for the other aspects of computing. 2) Integrating computer techniques with social science. We have to accept that in terms of privacy study, computing techniques and strategies have to follow or serve the needs and findings of social science, which is the leading battle ground. This is supported by the leading researchers in the field, such as the authors of the highly cited survey paper [11], and the emerging discipline of Computational Psychophysiology [9]. 3) Inventing new theoretical privacy frameworks. We have seen the practical usage of the various data clustering methods in privacy protection, and also the strictness of the differential privacy framework.

IV. RECENT TECHNIQUES OF PRIVACY PRESERVING IN BIG DATA

Differential privacy

Differential Privacy [10] is a technology that provides researchers and database analysts a facility to obtain the useful information from the databases that contain personal data of people without revealing the individual identities. This is done by introducing a minimum distraction in the information provided by the database organisation. The interference introduced is large enough so that they safeguard the privacy and at the same time small enough so that the information provided to analyst is still useful. Earlier some methods have been used to safeguard the privacy, but proved to be unsuccessful.

In mid-90s when the Commonwealth of Massachusetts Group Insurance Commission (GIC) released the anonymous health record of its clients for research to benefit the society [11]. GIC hides some information like name, street address etc. so as to protect their privacy. Latanya Sweeney (then a MIT, PhD student) using the freely available voter database and record released by GIC, successfully recognized the health record by just associating them. Thus hiding some data cannot assure the protection of distinct identity.

Differential Privacy (DP) deals to offer the solution to this problem as shown Fig. 2. In DP analyst is not providing the direct contact to the database containing personal data. An intermediate piece of software is hosted between the database and the analyst to protect the privacy. This intermediate software is also called as the privacy protector.

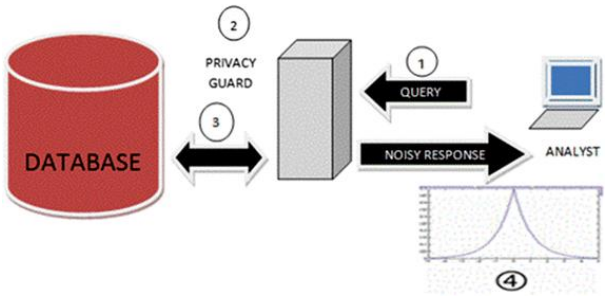

Fig. 2 Differential privacy as a solution to privacy-preserving in big data is shown

Step 1 The analyst can type a query to the database through this intermediate privacy guard.

Step 2 The privacy guard precedes the query from the analyst and weighs this query and other earlier inquiries for the privacy risk.

Step 3 The privacy guard then develops the answer from the database.

ISSN (O) 2454-9665
ISSN (P) 2454-0663

Step 4 Add certain distortion to it giving to the evaluated privacy risk and finally provide it to the analyst.

The quantity of distortion added to the pure information is proportional to the estimated privacy risk. If the privacy risk is small, distortion added is small enough so that it do not disturb the quality of answer, but big enough that they protect the specific privacy of database. But if the privacy risk is high then more distortion is added.

Identity based anonymization

These techniques faced issues when successfully shared anonymization, privacy protection, and big data techniques [12] to analyse practice data while protecting the characteristics of users. Intel Human Factors Engineering group wanted to use web page access logs and big data tools to improve convenience of Intel's deeply used internal web portal. To safeguard Intel employees' privacy, they were essential to remove personally identifying information (PII) from the portal's usage log repository but in a way that did not influence the application of big data tools to do examination or the ability to re-identify a log record in order to examine unusual behaviour. Cloud computing is a sort of large-scale distributed computing models which has become a driving potency for Information and Communications Technology over the past some years, due to its inventive and promising vision. It offers the possibility of educating IT systems management and is altering the way in which hardware and software are designed, purchased, and utilized. Cloud storage service brings important benefits to data owners (1) dropping cloud users problem of storage management and equipment maintenance, (2) evading investing a huge amount of hardware and software, (3) permitting the data access independent of geographical location, (4) retrieving data at any time and from anywhere [13].

To meet these objectives, Intel formed an open architecture for anonymization [12] that allowed a diversity of tools to be utilized for both de-identifying and re-identifying web log files. In the process of implementing architecture, establish that enterprise data has properties diverse from the standard examples in anonymization literature [13]. This idea showed that big data techniques might yield profit in the enterprise location even when operational on anonymized data. Intel also originated that despite masking clear Personal Identification Information like usernames and IP addresses, the anonymized data was powerless against correlation attacks. They discover the trade-offs of correcting these vulnerabilities and establish that User Agent (Browser/OS) information strongly correlates to individual users. This is a case study of anonymization implementation in an activity, describing requests, implementation, and experiences encountered when utilizing anonymization to guard privacy in enterprise data analysed using big data techniques. This examination of the quality of anonymization used k-anonymity based metrics. Intel used Hadoop to analyse the anonymized data and acquire valuable results for the Human Factors analysts [14,15]. At the same time, learned that anonymization needs to be more than simply masking or generalizing certain fields—anonymized datasets need to be carefully analysed to determine whether they are exposed to attack.

Privacy preserving Apriori algorithm using MapReduce framework

Hiding a needle in a haystack [16]

Existing privacy-preserving association rule algorithms alter original transaction data through the noise addition. However, this effort maintained the original transaction in the noised transaction in light of the fact that the goal is to avoid data utility deterioration while prevention the privacy destruction. Therefore, the opportunity that an untrusted cloud service provider infers the real frequent item set remains in the method [17]. Despite the risk of

ISSN (O) 2454-9665
ISSN (P) 2454-0663

association rule leakage, provide enough privacy protection because this privacy-preserving algorithm is based on "hiding a needle in a haystack" [16] concept. This concept is based on the idea that detecting a rare class of data, such as the needles, is hard to find in a haystack, such as a large size of data, as shown in Fig. 3. Existing techniques [18] cannot add noise haphazardly because of the need to consider privacy-data utility trade-off. Instead, this technique incurs additional computation cost in adding noise that will make the "haystack" to hide the "needle." Therefore, ought to consider a trade-off between problems would be easier to resolve with the use of the Hadoop framework in a cloud environment. In Fig. 3, the dark diamond dots are original association rule and the empty circles are noised association rule. Original rules are hard to be revealed because there are too many noised association rules [16]

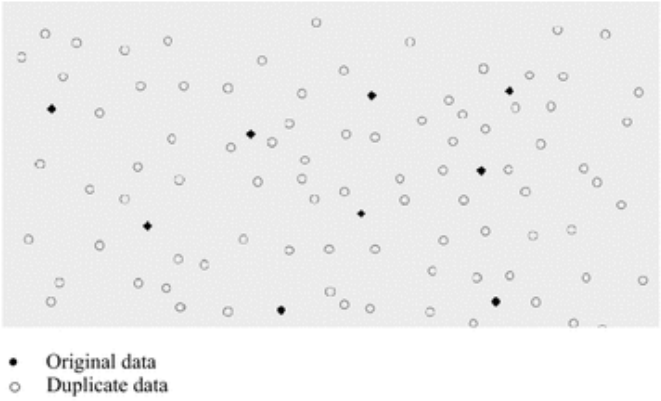

Fig. 3 Hiding a needle in a haystack Mechanism of hiding a needle in a haystack is shown

In Fig. 4, the service providers add a dummy item as noise to the unique transaction data collected by the data provider. Subsequently, a unique code is assigned to the dummy and the original items. The service provider preserve the code information to sort out the dummy item after the removal of frequent item set by an outside cloud platform. Apriori algorithm is performed by the external cloud platform using data which is sent by the service provider. The external cloud platform proceeds the frequent item set and support charge to the service provider. The service provider filters the frequent item set that is affected by the dummy item using a code to extract the correct association rule using frequent item set without the dummy item. The process of extraction association rule is not a burden to the service provider, considering that the amount of calculation required for extracting the association rule is not much.

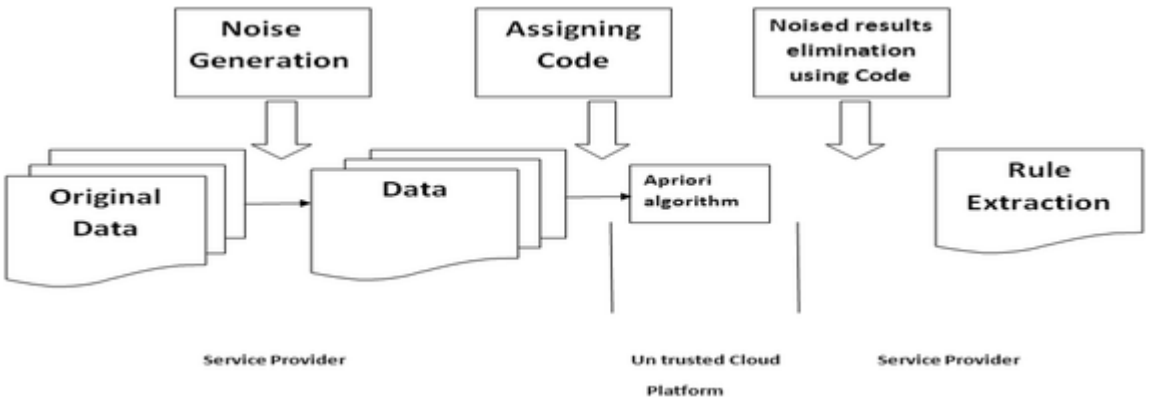

Fig. 4 Overview of the process of association rule mining the service provider adds a dummy item as noise to the original transaction data collected by the data provider

ISSN (O) 2454-9665
ISSN (P) 2454-0663

Privacy-preserving big data publishing

The publication and dissemination of raw data are crucial components in commercial, academic, and medical applications with an increasing number of open platforms, such as social networks and mobile devices from which data might be gathered, the volume of such data has also increased over time [19]. Privacy-preserving models broadly fall into two different settings, which are referred to as input and output privacy. In input privacy, the main concern is publishing anonymized data with models such as k-anonymity and l-diversity. In output privacy, usually interest is in problems such as query auditing and association rule hiding where the output of dissimilar data mining algorithms is perturbed or audited in order to preserve privacy. Much of the work in privacy has been dedicated on the quality of privacy preservation (vulnerability quantification) and the utility of the published data. The solution is to just divide the data into smaller parts (fragments) and anonymize each part independently [20].

In spite of the fact that k-anonymity can avoid identity attacks, it fails to guard from attribute disclosure attacks because of the need of diversity in the sensitive attribute within the equivalence class. The l-diversity model mandates that each equivalence class must have at least l well-represented sensitive values. It is common for large data sets to be processed with distributed platforms such as the MapReduce framework [21, 22] in order to distribute a costly process among multiple nodes and accomplish considerable performance improvement. Therefore, in order to resolve the inefficiency, improvements of privacy models are introduced.

Trust assessment plays an significant role in trust management. It is a technical approach of representing trust for digital processing, in which the factors influencing trust are evaluated based on evidence data to get a continuous or discrete number, referred to as a trust value. It suggests two schemes to preserve privacy in trust assessment. To reduce the communication and computation costs, propose to introduce two servers to realize the privacy preservation and evaluation result sharing among various requestors. Consider a scenario with two independent service parties that do not collude with each other due to their business incentives. One is an authorized proxy (AP) that is responsible for access control and management of aggregated evidence to enhance the privacy of entities being evaluated.

The additional is an evaluation party (EP) (e.g., offered by a cloud service provider) that practices the data collected from an amount of trust evidence providers. The EP processes the collected data in an encrypted form and produces an encrypted trust pre-evaluation result. When a user needs the pre-evaluation result from EP, the EP first orders the user's access eligibility with AP. If the check is positive, the AP re-encrypts the pre evaluation result that can be decrypted by the requester (Scheme 1) or there is an additional step involving the EP that prevents the AP from obtaining the plain pre-evaluation result while still allowing decryption of the pre-evaluation result by the requester (Scheme 2) [23].

V. CONCLUSION

The amount of data is growing everyday and it is impossible to imagine the next generation applications without producing and executing data driven algorithms. In this paper, we have conducted a comprehensive survey on the privacy issues when dealing with big data. We have investigated privacy challenges in each phase of big data life cycle and discussed about the recent techniques of privacy preserving in big data. Some of the future

ISSN (O) 2454-9665
ISSN (P) 2454-0663

research directions for big data privacy are access control and secure end to end communication, data anonymization, decentralized storage and effective machine learning techniques and distributed data analytics.

REFERENCES

[1] J. Manyika et al., Big data: The Next Frontier for Innovation, Competition, and Productivity. Zürich, Switzerland: McKinsey Global Inst., Jun. 2011, pp. 1–137.

[2] B. Matturdi, X. Zhou, S. Li, and F. Lin, ‘‘Big data security and privacy: A review,’’ China Commun., vol. 11, no. 14, pp. 135–145, Apr. 2014.

[3] J. Gantz and D. Reinsel, ‘‘Extracting value from chaos,’’ in Proc. IDC IView, Jun. 2011, pp. 1–12.

[4] A. Katal, M. Wazid, and R. H. Goudar, ‘‘Big data: Issues, challenges, tools and good practices,’’ in Proc. IEEE Int. Conf. Contemp. Comput., Aug. 2013, pp. 404–409.

[5] L. Xu, C. Jiang, J. Wang, J. Yuan, and Y. Ren, ‘‘Information security in big data: Privacy and data mining,’’ in IEEE Access, vol. 2, pp. 1149–1176, Oct. 2014.

[6] H. Hu, Y. Wen, T.-S. Chua, and X. Li, ‘‘Toward scalable systems for big data analytics: A technology tutorial,’’ IEEE Access, vol. 2, pp. 652–687, Jul. 2014

[7] D. Gottesman and I. L. Chuang, ‘‘Demonstrating the viability of universal quantum computation using teleportation and single-qubit operations,’’ Nature, vol. 402, no. 6760, pp. 390–393, 1999.

[8] S. Barz, E. Kashefi, A. Broadbent, J. F. Fitzsimons, A. Zeilinger, and P. Walther, ‘‘Demonstration of blind quantum computing,’’ Science, vol. 335, no. 6066, pp. 303–308, 2012.

[9] Advances in Computational Psychophysiology, accessed on May 17, 2016. [Online]. Available: http://www.sciencemag.org/custompublishing/collections/advances-computational-psychophysiology

[10] Microsoft differential privacy for everyone, [online]. 2015. http://download.microsoft.com/…/Differential_Privacy_for_Everyone.pdf.

[11] Sedayao J, Bhardwaj R. Making big data, privacy, and anonymization work together in the enterprise: experiences and issues. Big Data Congress; 2014.

[12] Yong Yu, et al. Cloud data integrity checking with an identity-based auditing mechanism from RSA. Future Gener Comp Syst. 2016;62:85–91.

[13] Oracle Big Data for the Enterprise, 2012. [online]. http://www.oracle.com/ca-en/technoloqies/biq-doto.

[14] Hadoop Tutorials. 2012. https://developer.yahoo.com/hadoop/tutorial.

[15] Fair Scheduler Guide. 2013. http://hadoop.apache.org/docs/r0.20.2/fair_scheduler.html

[16] Jung K, Park S, Park S. Hiding a needle in a haystack: privacy preserving Apriori algorithm in MapReduce framework PSBD’14, Shanghai; 2014. p. 11–17.

[17] Ateniese G, Johns RB, Curtmola R, Herring J, Kissner L, Peterson Z, Song D. Provable data possession at untrusted stores. In: Proc. of int. conf. of ACM on computer and communications security. 2007. p. 598–609.

[18] Verma A, Cherkasova L, Campbell RH. Play it again, SimMR!. In: Proc. IEEE Int’l conf. cluster computing (Cluster’11); 2011.

[19] Feng Z, et al. TRAC: Truthful auction for location-aware collaborative sensing in mobile crowd sourcing INFOCOM. Piscataway: IEEE; 2014. p. 1231–39.

ISSN (O) 2454-9665
ISSN (P) 2454-0663

[20] HessamZakerdah CC, Aggarwal KB. Privacy-preserving big data publishing. La Jolla: ACM; 2015.

[21] Dean J, Ghemawat S. Map reduce: simplied data processing on large clusters. OSDI; 2004.

[22] Lammel R. Google's MapReduce programming model-revisited. Sci Comput Progr. 2008;70(1):1–30.

[23] Yan Z, et al. Two schemes of privacy-preserving trust evaluation. Future Gener Comp Syst. 2016;62:175–89.