International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

COPY RIGHT

ELSEVIER
SSRN

IJIEMR Transactions, online available on 17th Apr 2019. Link
:http://www.ijiemr.org/downloads.php?vol=Volume-08&issue=ISSUE-04

Title: A SURVEY ON DATA PRIVACY AND SECURITY TECHNIQUES IN DATA MINING, CLOUD COMPUTING AND BIG DATA

Volume 08, Issue 04, Pages: 269–278.

Paper Authors

K SANDHYA RANI KUNDRA, PROF.P.V.G.D.PRASAD REDDY, PROF.K.VENKATA RAO

USE THIS BARCODE TO ACCESS YOUR ONLINE PAPER

To Secure Your Paper As Per UGC Guidelines We Are Providing A Electronic Bar Code

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

A SURVEY ON DATA PRIVACY AND SECURITY TECHNIQUES IN DATA MINING, CLOUD COMPUTING AND BIG DATA

[1]K SANDHYA RANI KUNDRA, [2]PROF.P.V.G.D.PRASAD REDDY, [3]PROF.K.VENKATA RAO

[1]Asst. Professor, Dept of I .T, G.V.P.Collage of Engg(A), Visakhapatnam

[2]Sr. Professor, Dept. of CS & SE, Andhra University, Visakhapatnam

[3]Professor, Dept. of CS & SE, Andhra University, Visakhapatnam

[1]sandhyaranikks38@gmail.com, [2]prasadreddy.vizag@gmail.com

Abstract—In an era of fast paced technological advancements, issues related to control and use of personal data are taking centre Stage. Personal data is often referred to as "the new oil of the Internet and the new currency of the digital world". The investigation of such information is encouraged businesses and development to the society in many different fields. However, the storage and flow of sensitive data bearing serious privacy concerns. While safeguarding privacy, techniques that let the knowledge extraction from data are known as privacy-preserving data mining (PPDM) techniques. Data security and privacy has persistently been a major issue in information technology. Privacy and security are constantly plays an important role in the online world. It is extensively accepted that cloud computing has the potential to be privacy debilitate. The secure processing of personal information in the cloud represents a big challenge. The day to day of digitalization activities has resulted in a large volume of data. This type of data, called Big Data, is used by many standardized companies to extract valuable information to take marketing decisions, as well as text analytics and sensor data or detect threat attacks. This paper surveys the most relevant PPDM methods from the literature and presents typical applications of PPDM techniques in relevant areas. Various issues related to data stored in cloud storage and solutions were listed up. The benefits of Big Data Analytics were emphasized first and then the challenges of security and privacy in big data environments were analyzed. Moreover, some protection techniques and some possible tracks that approve security and privacy in a pernicious big data context were also proposed. As a whole this paper provides a good summary and an overview of PPDM, cloud computing and big data.

Keywords— privacy, Security, privacy-preserving data mining, cloud computing, Big data.

I.INTRODUCTION

In current scenario we now spend much of our lives online, and our online activities-from shopping to socializing, entertainment to information Searching and gathering - have created an unprecedented number of data points. But these data bring risks. When governments opt to use these data for surveillance purposes, even when claimed to be for the public good, such data points can be used against us. On the other hand, this technological arrival also introduces novel ways of information leakage and user

classified data security and privacy issues over data is stored and transmitted over the cloud and even across borders. This seems very threatening to the cloud user's community, and they have raised very serious concerns about these issues. Although everyone has a concept of privacy and security, there is no universally accepted standard definition. Data privacy is suitably defined as the appropriate use of data. Privacy and security of these collected records are very important and high priority. Therefore, the research community has to consider these concerns by proposing and implementing strong protection mechanisms that permit from data without risking security and privacy. Data security is commonly referred to as the confidentiality, availability, and integrity of data. In other words, data isn't being used or accessed by unauthorized individuals or parties. This work will focus on the most important aspects of privacy preserving techniques in data mining, cloud computing and big data on privacy and security. In the information field, privacy as ``the right of an individual to be secure from unauthorized disclosure of information about oneself that is contained in an electronic repository". or in other words, as the right to control the handling of one's information. Bertino et al. [1] gave a similar definition. The Second section discusses the most important challenges in privacy preserving data. In Third section we describe various service and deployment models of cloud computing and identify major challenges. The fourth section we discuss some related work to secure the big data concepts and also propose few possible solutions that approve a privacy protection in a big data context. Finally, Section V concludes some future aspects towards a secure big data, PPDM and cloud computing rising areas.

II. DATA PRIVACY IN DATA MINING

Privacy Preservation in data mining has crop up as an outright essential for exchanging confidential information in terms of analysis of data, data validation, and data publishing. A few benefits of the information technologies are only possible through the collection and analysis of sensitive data. For that, this may result in unwanted privacy violations. To protect from information leakage, privacy preservation methods have been developed to protect owner's disclosure, by modifying the original data [2], [3]. However, transforming the data may also reduce its efficiency, resulting in faulty or even infeasible extraction of knowledge through data mining. This scenario is known as Privacy-Preserving Data Mining (PPDM). PPDM methodologies are designed to a convinced level of privacy, while maximizing the utility of the data, so that data mining can still be performed on the reconstruct data efficiency. PPDM encompasses all techniques that can be used to extract knowledge from data while preserving privacy. This chapter provides an overview on new perspective and systematic interpretation of a list published literatures via their meticulous organization in subcategories. The fundamental notions of

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

the existing privacy preserving data mining techniques, their merits are presented in Table-1. The current privacy preserving data mining techniques are classified based on distortion, association rule, hide association rule, taxonomy, clustering, associative classification, outsourced data mining, distributed, and k-anonymity, where their notable advantages and disadvantages are emphasized. This PPDM may consist on using data transformation techniques, such as the ones in Table-2. . PPDM also used for the distributed privacy techniques [20] that are employed for mining global insights from distributed data without disclosure of local information. Due to the variety of proposed techniques, Several metrics to evaluate the privacy level and the data quality or utility of the different techniques have been proposed [1],[4],[5]and[6]. This careful scrutiny reveals the past development, present research challenges, future trends. Further significant improvements for more powerful privacy protection and preservation are confirm to be mandatory.

TABLE 1: *Summary of privacy-preserving techniques at data publishing in terms of sanitization methods.*

Privacy Models	Description	Advantages
k-anonymity [14][15](Generalization, Suppression)	Anonymity is guaranteed by the existence of at least other k-1 undistinguishable (w.r.t the QID) records for each record in a data base.	1. Simplicity of definition. 2. Great amount of exiting Algorithms.
l-diversity[16] (Generalization, Suppression)	Expands the k-anonymity model by requiring every equivalence class to have at least l"well represented" value for sensitive attributes.	The diversity of sensitive attribute values taken into consideration for anonymization
t-closeness[17] (Generalization, Suppression)	Solve l-diversity problem of skewed sensitive values distribution by requiring that the distribution of sensitive values in each equivalence class "close" to the corresponding distribution of original table, where close means upper bounded by the threshold value.	Takes into consideration the distribution of the sensitive values when forming the equivalence class.
Personalized Privacy[18] (Generalization)	Achieved by creating a taxonomy tree using generalization by allowing the record owners to define guarding node. Owner's privacy breached if an attacker allowed inferring any sensitive value from subtree of guarding node with a probability (breach probability) greater than certain threshold.	1. Owner's can define their privacy level. 2. Preserves maximum utility while respecting personal privacy preferences.
€-differential privacy Perturbation[19] (Randomization)	Ensure that a single record does not considerably affect the outcome of the analysis of the dataset. In this sense, a person's privacy will not be affected by participating in the data collection since it will not make significant difference in the final outcome.	Provides a formal privacy guarantee and a solid privacy loss metric.

TABLE-2: *Summary of the privacy-preserving techniques at data collection.*

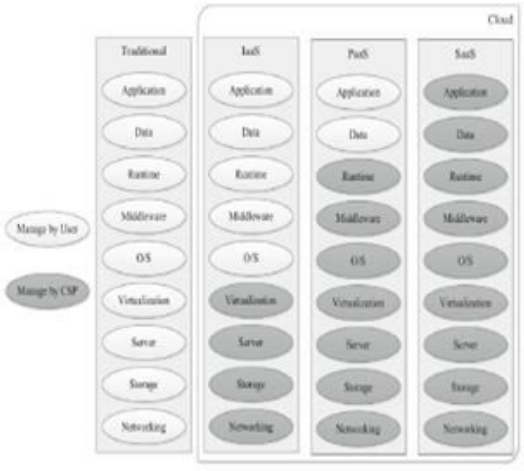
Randomization Method	Description	Advantages
Additive Noise[21]	Data is randomized by adding noise with a known statically distribution.	1. Performs independently for each captured value. 2.Preserves statically properties after reconstruction of the original distribution
Multiplicative Noise[22]	Data is randomized by multiplying noise with a known statistical distribution.	1. More effective than additive noise at preserving privacy, since the reconstruction of the original individual values is more difficult. 2.Performs independently for each captured value(suitable for data collection)

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

These presentations highlight the significant development of privacy preserving data mining methods and fundamental observation. Several perspectives and new clarifications on privacy preserving data mining approaches are accomplished. Existing literatures are systematically subcategorized to identify the strengths, gap, and weakness of various approaches. This section primarily focused on the creation of awareness and relevant action to be taken by all relevant techniques to protect privacy in secured data transfer over the web.

III. DATA PRIVACY IN CLOUD COMPUTING

Cloud computing refers to applications and services offered over the Internet. These services are offered from data centers all over the world, which collectively are referred to as the "cloud". Data security and privacy protection are becoming more important for the future development of cloud computing technology in all organizations. Data security and privacy protection issues are relevant to both hardware and software in the cloud architecture.The explanation of "cloud computing" from the National Institute of Standards and Technology (NIST)[7] .Cloud computing is an Internet based computing which enables sharing of services. Using Cloud Storage, users can store their data remotely and enjoy the on-demand high quality applications and services from a shred pool of configurable computing resources. The cloud storage has many benefits over local data storage. User should be able to just use the cloud storage as if it is local, without worrying about the need to verify its integrity. Cloud provides many services like in online marketing, banking and payment, health care centers, social media as per use of personal information. Those privacy-sensitive data stores other side of the globe. Privacy and security in cloud just like how privacy of users is protected and observed.Cloud computing can be identified as a new computing paradigm that can provide services on demand at a minimal cost. The three well-known and commonly used service models in the cloud are software as a service (SaaS), platform as a service (PaaS), and infrastructure as a service (IaaS). In the below figure (Service Oriented cloud computing Architecture [12]) shows the comparison between those models with the traditional model.



SaaS software, it allows people to access the functionality of particular software without worrying about storage or other issues.

PaaS a service provider, companies can run their applications on the cloud service's platform without having to worry about maintains hard drivers and servers.

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

In IaaS, Organizations make use of unlimited storage potential of the cloud infrastructure. Because of the dedicated server on site they have the leverage to expand or shrink the storage space.

According to the difference of access scope, cloud can be divided into three types: public cloud, private cloud, and hybrid cloud. In "Public Cloud" the cloud is used by general public. It is owned and operated by a single user/client. In "Private Cloud", the cloud infrastructure is access by a single organization or IT sector. "Hybrid cloud" is a mixture of two deployment models (private or public). Its compositions are liable to allow a technology that authorize data and application portability. Most of the existing cloud services are provided by large cloud service companies such as Google, Amazon, and IBM.There are three major potential threats in cloud computing, namely, security, privacy, and trust.

A data security framework for cloud computing networks is proposed [8]. The primary challenge in cloud computing is data sharing and data security issues at SPI (SaaS, PaaS, and IaaS).In this chapter, we will review different security techniques and challenges for data storage security and privacy protection in the cloud computing environment. As Figure 1 shows, this chapter presents a comparative research analysis of the existing research work regarding the techniques used in the cloud computing through data security aspects including data integrity, confidentiality, and availability. In the cloud computing environment securing data to enhance the users trust on comparative studies on data security and privacy

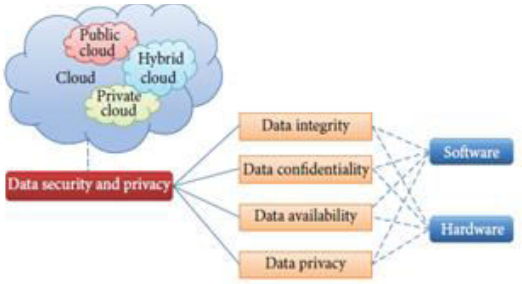


FIGURE I: Organization of data security and privacy in cloud computing.

In the cloud, the privacy means when users visit the sensitive data. The privacy issues differ according to different cloud scenarios and can be divided into four subcategories [9, 10, 11] as follows:

• How to enable users to have control over their data when the data are stored and processed in cloud and avoid theft, nefarious use, and unauthorized resale,

• how to guarantee data replications in a jurisdiction and consistent state, where replicating user data to multiple suitable locations is an usual choice, and avoid data loss, leakage, and unauthorized modification or fabrication, who is responsible for the personal information and the legal requirements that encompass it.

• To what extent cloud subcontractors are involved in processing which can be properly identified, checked, and ascertained.Issues in cloud storage about Privacy and security:

• The treats against information assets residing in cloud computing environments.

•The capacity to attack the cloud and its types.

• The security risks associated with the cloud, and where relevant considerations of attacks and counter measures.

• Emerging cloud security risks.

Solutions for various privacy preserving techniques:

I. Access control Mechanism: A privacy preserving authenticated access control scheme is used for security of user data in cloud storage. Authorized users can decrypt the stored data. It preserves the privacy of data; maintain the security and keeps secrete the identity of user.

II. Auditability schemes: To improve their services but also reduces the online burden. The auditing is of two types like public auditability and private Auditability. The private auditability gives higher efficiency. The public auditability enables everybody like users customers to interact with cloud server or cloud storage.

III. Public Verification for Storage Security: To maintain the data integrity the user may resort to TPA (Third Party Auditing).So the TPA performs the auditing on behalf of the user and reduces online burden.

IV. Data Integrity Checking for Privacy-Preserving: Is used for checking protocol. This Protocol provides public verifiability without help of a TPA. It doesn't disclose any user information to TPA.

These different techniques and methods are used to solve the problems of privacy of user data.

IV. DATA PRIVACY IN BIG DATA

Big data is a term used for massive amount of data sets that have more mixed and complex Structure. Big data carry with it new security concerns. Present days our data capacity is growing exponentially, we have inexact solutions for the many security issues that affect even local, self-supporting data.Big Data is used by many grouping Managements to extract useful information either to take marketing choices, track exact behaviors or detect menace attacks. The processing of data is made possible by using multiple techniques, called Big Data Analytics. It examines large amount of data to uncover hidden patterns, correlations and insights with any huge volume of unstructured, structured and semi-structured data.Risk arises from the analytics tools consist of storing, managing and efficiently analyzing various data gathered from all possible and available sources. It is possible to collect more data than it should have which leads to many security and privacy issues. For this reason, Investigation center has to consider these issues by proposing strong protection techniques to big data without imperil privacy. In this chapter, we highlight the benefits of Big Data Analytics and also review the challenges of security and privacy in big data ambiance.

Challenges in Big Data:

1. Privacy and Security: Is the most important challenges with Big Data which is sensitive and legal significance.

• The personal information of an individual when combined with external large data sets, leads to inference of new facts about that individual.

• Information regarding the individual is collected and used in order to add worth to the business organization. This could be

achieved by creating insights in their lives which they are ignorant of.

- Law enforcement body by using Big data increases the probability of people to suffer from dire consequences even without the knowledge of those tagged people that they are discriminated and without the ability to fight back.

2. Data Access and Sharing of Information: Expecting sharing of Data between companies is roused cause of the need to get an edge in business. The secrecy and competitiveness is threatened by sharing the client's data.

3. Data Analysis: This type of analysis to be done in on this huge amount of data which can be either unstructured, semi- structured or structured requires a large number of advance skills. The results that will be obtained decides the type of analysis needed to be done on the data i.e. decision making

4. Various sources of Data: Dealing with the large volume Of data being produced s a challenge .Additionally, it is challenge to manage the enormous number of sources that are producing this data. The data comes from companies internal sources like finance, marketing etc.

5. Data Storage and Retrieval: Current available technologies are able to handle data entry and data storage. But the tools designed for transaction processing which will add, update, and search for small to huge amount of data is unable to handle big data. By what means to handle semi or unstructured data for processing it is not known [23].

6. Data Growth and Expansion: As the organizations increases their services, their data are also expected to grow. Some organizations also consider data expansion because of data grow in richness and data evolved with new trends and technologies [23].

7. Speed and Scale: Whenever new volume of data grows, it is difficult to gain insight into data within time period. Gaining insight into data is more important than processing complete set of data. Processing near real time data will always require processing interval in order to produce satisfactory output [23].

Techniques to Protect Privacy in Big Data:

The big data Environment is the composition of several technological evolutions for both storage and processing capabilities. For this reason traditional security techniques cannot be efficient and directly applied to big data contexts. Therefore, new security techniques should be deployed to accompany these evolutions. The big challenge when adding security and privacy to big data platform is to come up with balanced solutions between regulations, security controls, and analytics. Before data set is out for other parties, the possibility of identifying sensitive information about individuals is reduced by some privacy-preserving technique. This is called the disclosure-control problem [24][25].There are some approaches to preserve privacy:

1. Rule based access control: Adding Certification or access control to the data entries restricts the data so that limited

groups can access sensitive information. Challenges here is that no sensitive information can be misconduct by unauthorized individuals and thus secured certification or access control mechanisms must be designed.

2. Cryptographic Technique: sensitive information fields should be anonymized so that they cannot pinpoint to an individual record. The main challenge is to insert randomness into the data to ensure a number of privacy goals [38].

3. Data Anonymization: Anonymization of data is another possible way to protect collected big data. It is also referred as data de identification. The basic idea consists of using data perturbation and data swapping techniques to protect the association of individuals to critical information. The annonymization includes following approaches:

i. Generalization & Suppression- This is generally used to replace the specific values with more general ones that leads to many tuples will having duplicate values for quasi identifiers. We replace quasi identifiers by some constant vales like 0,* etc.

ii. Anatomization & Permutation-This is de-linking between quasi-identifiers and sensitive attributes.

iii. Perturbation- The process of adding some noise to the original data before giving the data to the user.

There are mainly three privacy-preserving methods based on data anonymization are K-Anonymity [13], L-Diversity [16] and T-Closeness [17].

4. Differential Privacy: Is a method enabling analysts to extract useful answers from databases containing personal information while offering strong individual privacy protections [26].It aims to minimize the chances of individual identification while querying the data.

Differential privacy has some advantages over anonymization. The advantages of differential privacy are:

i. The original data set is not modified at all. There is no need for suppression or generalization.

ii. Distortion is added to the results by mathematical calculations based on the type of data, type of questions etc.

iii. The distortion is added in such a way that value hidden is useful to analysts.

V.CONCLUSION

A large number of cloud datacenters are established and they provide necessary tools and infrastructure to utilize economical, storage services, on-demand and rapid-elasticity computation. To many innovation and marketing strategies, analytics and mining have been regarded as the key enabler in unlocking the value of bigdata, which has pressed more efforts to the bigdata related to R&D. As a proof for instance, Gartner has reported that most of the world's largest 200 companies have planned to invest in the development of intelligent apps and to utilize analytics tools by 2018 for the full toolkit of bigdata. New founding from these investments is to be incorporated to refine the customer experiences through services offered by companies. This explains that a broad

research is expected and to be supported more actively for big data mining infrastructure, platforms, and applications that runs on wired and wireless communication channels to conduct more efficient knowledge smart decision support and discovery.

REFERENCES

[1] E. Bertino, D. Lin, and W. Jiang, ``A survey of quanti_cation of privacy preserving data mining algorithms," in Privacy-Preserving Data Mining. New York, NY, USA: Springer, 2008, pp. 183_205.

[2] C. C. Aggarwal and P. S. Yu, ``A general survey of privacy-preserving data mining models and algorithms," in Privacy-Preserving Data Mining. New York, NY, USA: Springer, 2008, pp. 11_52.

[3] C. C. Aggarwal, Data Mining: The Textbook. New York, NY, USA:Springer, 2015.

[4] S. Dua and X. Du, Data Mining and Machine Learning in Cybersecurity.Boca Raton, FL, USA: CRC Press, 2011.

[5] E. Bertino and I. N. Fovino, ``Information driven evaluation of data hiding algorithms," in Proc. Int. Conf. Data Warehousing Knowl. Discovery,2005, pp. 418_427.

[6] S. Fletcher and M. Z. Islam, ``Measuring information quality for privacy preserving data mining," Int. J. Comput. Theory Eng., vol. 7, no. 1,pp. 21_28, 2015.

[7] P.Mell and T. Grance, "The nist definition of cloud computing,"National Institute of Standards and Technology, vol. 53, no. 6, article 50, 2009.

[8] A. Pandey, R. M. Tugnayat, and A. K. Tiwari, "Data Security Framework for Cloud Computing Networks," International Journal of Computer Engineering & Technology, vol. 4, no. 1, pp.178–181, 2013.

[9] S. Pearson and A. Benameur, "Privacy, security and trust issues arising from cloud computing," in Proceedings of the 2nd IEEE International Conference on Cloud Computing Technology and Science (CloudCom '10), pp. 693–702, IEEE, December 2010.

[10] S. Paquette, P. T. Jaeger, and S. C. Wilson, "Identifying the security risks associated with governmental use of cloud computing," Government Information Quarterly, vol. 27, no. 3, pp.245–253, 2010.

[11] S. Subashini and V. Kavitha, "A survey on security issues in service delivery models of cloud," Journal of Network and Computer Applications, vol. 34, no. 1, pp. 1–11, 2011.

[12] M. Sookhak, H. Talebian, E. Ahmed, A. Gani, and M. K. Khan, "A review on remote data auditing in single cloud server: Taxonomy and open issues," Journal of Network and Computer Applications, vol. 43, pp. 121–141, 2014.

[13]. L. Sweeney, "k-Anonymity: A model for protecting privacy, " In International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, Vol. 10, No. 5, pp. 557-570, 2002.

[14]. P. Samarati and L. Sweeney, ``Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and

suppression," in Proc. IEEE Symp. Res. Secur. Privacy, 1998, pp. 384_393.

[15] P. Samarati and L. Sweeney, ``Generalizing data to provide anonymity when disclosing information," in Proc. PODS, 1998, p. 188.

[16]. A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "-diversity: Privacy beyond k-anonymity," ACM Trans. Knowl. Discovery Data, vol. 1, no. 1, p. 3, 2007.

[17]. N. Li, T. Li, and S. Venkatasubramanian, ``t-closeness: Privacy beyond k-anonymity and l-diversity," in Proc. IEEE 23rd Int. Conf. Data Eng. (ICDE), Apr. 2007, pp. 106_115.

[18]. X. Xiao and Y. Tao, ``Personalized privacy preservation," in Proc. VLDB, 2006, pp. 139_150.

[19]. X. Xiao and Y. Tao, ``Personalized privacy preservation," in Proc. VLDB, 2006, pp. 139_150.

[20]. C. Clifton, M. Kantarcioglu, J. Vaidya, X. Lin, and M. Y. Zhu, ``Tools for privacy preserving distributed data mining," ACM SIGKDD Explorations News lett., vol. 4, no. 2, pp. 28_34, 2002.

[21]. R. Agrawal and R. Srikant, ``Privacy-preserving data mining," ACM SIGMOD Rec., vol. 29, no. 2, pp. 439_450, 2000.

[22]. J. J. Kim and W. E. Winkler, ``Multiplicative noise for masking continuous data," Statist. Res. Division, U.S. Bureau Census, Washington, DC, USA, Tech. Rep. 2003-01, 2003.

[23].Stephen K, Frank A, J. Alberto E,William M,"Big Data :Issues and Challenges Moving Forward",IEEE, 46th Hawaii international conference on contemporary on System Sciences,2013.

[24] Data Mining With Big Data Xindong Wu, Fellow, IEEE, Xingquan Zhu, Senior Member, IEEE, Gong-Qing Wu, And Wei Ding, Senior Member, IEEE, IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 1, January 2014.

[25] Review on Data Mining with Big Data" Vitthal Yenkar, Prof.Mahip Bartere, IJCSMC, Vol. 3, Issue. 4, April 2014

[26] J. Salido, "Differential privacy for everyone," White Paper, Microsoft Coproration, 2012.