IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

AN OVERVIEW OF PRIVACY PRESERVATION IN BIG DATA

Kanika[*1], Alka Agrawal[2] & R.A. Khan[3]

[*1,2&3]Department of Information Technology, Babasaheb Bhimrao Ambedkar University, Lucknow-226025, U.P., India

DOI: 10.5281/zenodo.822970

ABSTRACT

Mobile with mega pixel cameras, social networking sites, wireless sensor networks, earth-orbiting satellites, finance market etc continuously generating more and more data than ever before: 90% of world's data has been produced in last two years. This huge amount of data is generated by different sources. Verification, authorization, authenticity, privacy such types of security issues is becoming major concern of big data in heterogeneous environments. To maintain the authenticity and preserve privacy of large amount of data is always an issue. In this paper authors discussed the data privacy and security issues such as node authentication, attribute linkage, degree attack distributed data, data loss etc. Authors also discussed some security techniques which help to maintain and preserve privacy and security of data. Main focus of this paper is to give an overview of the current issues in big data that can violate the individual's privacy.

KEYWORDS: Big Data, Its Architecture, Security and Privacy issues and its approaches.

I. INTRODUCTION

Huge amount of data is generated day by day which is collected from various sources such as satellite data, sensors data, social sites; stock market etc. this data is increasing day by day and comes in different formats like semi-structured, structured and unstructured which is messy and disorganized. This type of data cannot be handled, captured and processed by traditional techniques. Huge amount of data with different formats is known as 'big data' [1]. The velocity, at which innovative data are being created is staggering, data, is too huge, and runs too fast. Hence this data is so enormous which generate issues that how to analyze, store and process the data using existing software approaches and traditional database management system. To fetch the value from such type of data, a diverse technique is needed to process that data [8, 15]. Creation of data size per day is estimated as 2.5 Exabyte (1018). IBM specifies that 2.5 exabytes of data are created every day also in moreover last two years 90% of the total data has been produced. Data size from the several collecting sources has faced exponential growth, creating new technical and application challenges [25]. A universal project on big data was accomplished in 2012, which provided creation, processing, visualizing, storage, communication in real time world by collecting and analyzing large amounts of data. Today, Internet has become a vital Component of individual life and without internet life of individual is impossible [24].

In this heterogeneous environment, data is generating continuously and this large amount of data brings several security and privacy issues and the challenges such as authentication, data security data privacy. Security of user's data is the process of protecting and maintaining the confidentiality, integrity, and availability of data whether in storage, processing or communication. Table 1 describes the features of big data [32].

Table 1: Size of Big Data

Components	Big Data	Unit	Size
Model	No Schema	Terabyte	10^{12}
Formats	Unstructured, Semi-structured, Structured	Petabyte	10^{15}

Volume	Petabytes, Exabytes, Zettabytes, Yottabyte	Exabyte	10^{18}
Data Architecture	Distributed	Yottabyte	10^{24}

With this privacy of an Individual is still a critical problem that can only be handled with some serious solution in the area. Mostly users are worried about his/her data privacy because private information can be gathered by organizations and can be sell to the other parties and reused without the users consent [6]. Furthermore, in big data several challenges and threats, including privacy, confidentiality, integrity, and availability of data is exist, that have to be addressed [20]. The main objective of this study is to analysis the status of big data, characteristics, trends and new possibilities in big data technologies development, find out the privacy issues interconnected to the big data. The contribution of this paper proposed the big data architecture and designs the flow of big data in distributed environments presented. With this, authors discuss the main elements of architecture such as big data sources, its format: structured, semi-structured, etc., 4Vs of big data such as volume, variety, velocity, veracity and analysis [19].

The rest of the paper organized is as follows. Section two provides the literature review on big data and security. In section three, authors explain the architecture of big data. In section four types of attributes have been described. In section five various security and privacy issues have been explained. Section six is an analysis of security approaches used to secure big data. Finally, we conclude in section seven.

II. LITERATURE REVIEW

Vikram Phaneendra et.al [1], demonstrated that in past days the data was in as smaller amount, and RDBMS simply operated itbut today it is hard to handle vast data through RDBMS tools, which is knownas big data. They have described, that big data differs from other data in 5 dimensions such as volume, velocity, variety, value, and veracity. Albert Bifet et.al [14] stated that streaming data analysis in actual time is flattering the greatest and most disciplined way to acquire valuable knowledge, allowing organizations to respond quickly when problem emerge or identify to improve performance. The techniques used for big data are a storm, cascading, Apache Hadoop, Scribe, Apache big, etc.

Jonathan Stuart Wardet.al [15] did a survey on big data definition; two ideas mainly connect big data: data storage and data analysis. This, therefore, lifts the question that as to how big data is dissimilar from conventional data processing techniques. This short paper attempts to assemble the diverse definitions which have expanded the number of degrees of traction and to secure a clear and brief definition. Kalyani Shirudkar et.al [3] described that big data performing data operations and computation for huge amounts of data, remotely from the data owner's enterprise.They found some challenges such as scalable data mining and analytics, computation access control, and secure communication etc. They used diverse security methods like Type Based keyword search. Colin Tankard [21] have explained some recommends providing better control over big data sets, such as archiving, data leakage prevention and access control should be brought together. He has described that big data centralized storage is so sensitive. It creates new security challenges. Archana R.A et.al [22] described that big data is the extension of data mining. They explained MOBAT technique in this paper and proposed a data mining technique which secured original set of data in big data. Hakan Ozkose et.al [23] has described the process of big data. They explained that how big data comes in the picture? They gave the detail literature review of big data and explained, yesterday, today, and future of big data. They explained that storing and processing of data become difficult and classical approaches remain incapable of doing such transactions.

III. ARCHITECTURE OF BIG DATA

Big Data is not only a database; it assembled the core technologies and mechanisms for huge data analysis and data processing. It has multifarious components to collect, acquisition, and process, and to analyze for delivery results to target users and applications. The study shows that the major concern is simply to recognize the nature of big data, its characteristics, finding out the issues and problems associated with big data. In this research, authors have designed architecture of big data to defining the big data' existing challenges and security issues [19].The proposed architecture of big data contains the following six components: sources, collection, format of big data, it's characteristics, and analysis. These components address different aspects of big data. Big data is gathered from several sources such as social networking sites, etc. after the collection, data is acquired in one place where it is categorized such as Metadata, unstructured, structured, etc.

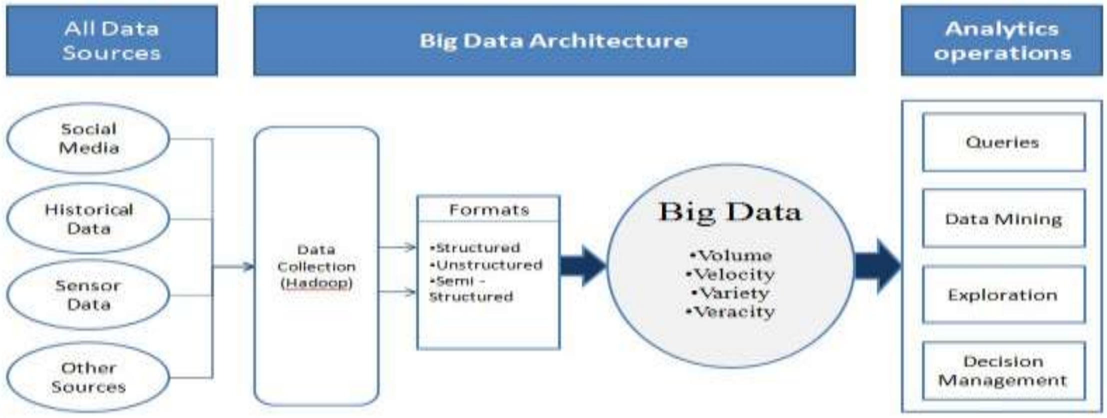


Figure 1: Architecture of Big Data

1. Sources

Data is created and collected from different sources. In figure 2 author shows some examples of data such as E-commerce, online transactions, etc. Nowadays on one has time to stand up in the queue; on one click they can keep update to yourself [25].

2. Format

Data is generated very speedily and comes in a various variety. Today's, data in various formats such as semi-structured, unstructured, hybrid,etc., Data is rising day to day in the volume and gathered by many organizations, such as social media, multimedia, and Internet of Things [25].

3. Characteristics of big data

Here described the characteristics of big data:

Volume

Volume means to the size of data which is generated day by day through user's internet activity. Now the size of data is larger than terabytes and petabytes [4, 5].

Velocity

Velocity refers to the creation speed of data or how quick the data is produced and meets the demands and the challenges which lie ahead of growth and development [3, 4].

Variety

Variety makes big data really big. It refers to the different types of data collected via B2B processes, logs, social media, streams, etc. Variety demotes to the structural diversity of elements [5].

Veracity

Due to a variety of data sources, intermediary processing, and in data growth raises a question about trust, privacy, security, and accountability, creating a need to verify protected data provenance [25].

Analysis

Big data analytics is the process of analyzing vast data sets to identify hidden patterns, hidden correlations, market patterns etc. Using advanced analytics techniques such as queries, data mining, exploration, statistics, and OLAP, etc, can analyze previously undiscovered data sources independent or together with their current enterprise data to gain new insights knowledge in significantly better and quicker decisions [34].

IV. SECURITY AND PRIVACY ISSUES

Big data is becoming a vital paradigm for real-time processing of enormous continuous data flows in huge heterogeneous networks. Security and privacy play an important role in the context of big data. Data security not only involves the encryption of the data but also ensures that appropriate policies are enforced for data sharing. The speedy data introduce huge open access and vast difficulties for big data [18].

Here some security issues such as degree attack, record linkage attacks, data mining based attacks, network level issues, authentication level issues, data level issues, and generic level issues etc has been discussed. This type of security issues occurs when there is a large volume of confidential data stored in a database, that is not encrypted or nor in the regular format [11].

Table 2: Privacy and Security Issues and its defenses

Privacy and security attacks	**Description**	**Suggested defense**
Phishing	Attempt to acquire sensitive information often for malicious reasons [31].	Security awareness program
Unauthorised access	When a person wants to access a website, services, or another system through else's account or other methods [33].	Access control
Degree Trial Attack	It re-identifies the nodes that belong to a target individual from a series of available graphs by comparison the degree of the nodes within the available graphs with the degree advancement of a target [26].	K-structural Diversity Anonymity
Data Mining Based Attacks	Using data mining methods to extract sensitive knowledge [31].	Divide datasets (vertically and Horizontally) and use access control.
Record Linkage Attack	Hacker is able to recognize the item of a target user by linking the item to data from different sources [29].	K-anonymity

To provide data protection in the presence of other unauthorized user is more difficult and when data is moving from homogeneous to the heterogeneous data, thus to handle and manage this data with the existing and traditional security tools and technologies is not capable. Here authors' discussed security issues in the context of authorization, authentication and data protection. They are as following:

A. Record linkage

Hacker is able to recognize the item of a target user by linking the item to data from different sources, for e.g. Liking the item to an item in a published data table. It is a difficult task because single entity identifiers are not accessible in every database that is connected. Thus, the common attributes available which are sufficiently well correlated with entities, known as quasi-identifiers (QI), have to be used for the linkage [26] [29].

B. Degree Attack

It re-identifies the nodes that belong to a target individual from a series of available graphs by comparison the degree of the nodes within the available graphs with the degree advancement of a target. The ability of this attack is that the attacker can actively manipulate the degree of the target individual by interacting with the social network [26].

C. Network level

Network level issues deal with network security and network protocols, such as Internodes communication, distributed nodes, distributed data. Many nodes are present in clusters andon those nodes, processing or computation of data is done. In a cluster, this processing of data can happen anywhere among the nodes. So it is challenging to discover the node, where data processing is happening. Because of this problem, providing security to processing node is going to be complicated [8, 16].

D. Structural Attack

Structural attack uses the additive degree of n-hop neighbors of a vertex because the regional feature, and join it with the imitation annealing-based graph matching methodology to re-identify vertices in anonymous social graphs [28].

E. Authentication Level

User authentication level challenges deal with encryption/decryption techniques, authentication methods (like administrative rights for nodes authentication of applications and nodes), and logging. Several nodes are present in a cluster. Each node has a different priorities or rights. If malicious nodes get the administrative rights, then it will steal or modify the confidential data. In the case of no authentication, affected node can destroy the cluster. In big data, Logging plays an essential job. If logging is not given, then no action is recorded. If a new node enters in the cluster then that will not be identified because of logging absence [8][10].

F. Data Level

Data level issues deal with data availability and integrity, such as data protection and distributed data. In a heterogeneous environment, information is stored in numerous nodes with replicas for fast retrieval. But if any copy or information is deleted or modified from another nodeby a hacker, then it will be challenging to recover data [8][10].

G. Generic Types

In distributed environment, various technologies used for data processing, as well as some traditional security tools for providing security features. Some traditional security tools were developed over years ago. So these tools may not be compatible with new distributed structure of big data. As big data uses several technologies for data retrieval, storing, and processing, certain complexities may occur because of the wide use of different technologies [8][10].

H. Distributed Data

To reduce parallel computation, a huge data set can be put away in several pieces across numerous machines. Additionally, redundant copies of data are made to guarantee the data reliability. If a particular chunk is corrupted, then data can be recovered from its copies. In the heterogeneous environment, it is extremely difficult to find the exact location, where of a file are stored. Also, these bits of information are replicated to another machines/node based on accessibility and maintenance operations [11].

V. TYPES OF ATTRIBUTES

In distributed environment user creates information. It can be sensitive or non-sensitive. This information is collected by the data collector/organization, so it is necessary for the data collector/organization to modify the actual data before releasing them to others to preserve its confidentiality. The actual information is assumed to be an individual table consist many records. Every record consists 4 types of attributes [26]:

- Identifier (I): Those Attributes that can uniquely identify a person, such as ID number, name.
- Quasi-Identifier (QI): Attributes that can be linked with external data to re-identify individual records, such as gender, DOB and pin code.
- Sensitive Attribute: Attributes that a user wants to hide, such as mobile and email id.
- Non-sensitive Attribute: other than Identifiers, Quasi-Identifier, and Sensitive Attribute.

The table should be anonymized before being transferred to others, i.e., identifiers are removed and quasi-identifiers can be modified. As a result, the identity of a person and values of a sensitive attribute can be concealed from the adversary [26].

VI. APPROACHES TO PRESERVE PRIVACY AND SECURITY IN BIG DATA

In today's scenario, people are dependent on the internet for most of the work, that's why data is growing large due to social media, banking sectors, online payments, etc. In this technology's era where everything is on fingertip such as online shopping, online banking, communication, etc. but everyone is not bothered about the data authenticity, its privacy, and security. User's privacy can be affected because of the unauthorized access to private data [24].

Here we show the figure 2 of Communication Process. In this figure, it is shown that how communication happens between the users and web server. How is it authorized? In this figure, there are two users, social domain networks, and a company. Users and social network upload and access data from the server, but a user can only access data when a user authorized to its [19]. This figure shows the communication between the user and server and how a company analyze user's data and sell it to the third party. A company access data through the server and sell it to the third party without user consent. The user is not aware such type of indirect attacks.

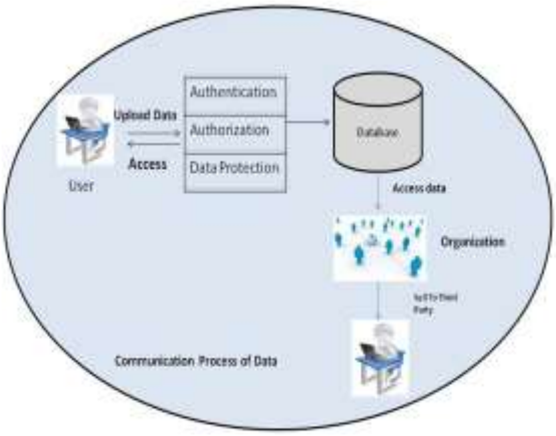


Figure 2: Communication Process of Big Data

In a heterogeneous environment, there are numerous third parties who may have chances to reach sensitive information. Nevertheless, there is no effective approach for preventing communicating data from the unauthorized, unauthenticated users and third parties [17].

A hacker can easily access or steal private information, misuse of this data. These are the basic techniques of data mining of issue overcome from these attacks, many solutions available such as k-anonymity, clustering, trajectory data, file encryption, honeypot, node authentication, access control etc [11, 26]. authors describe various solutions which collectively will make the environment secure. To deal with the privacy issues here researcher presents an overview of many security and privacy approaches which can help to improve the confidentiality and integrity of sensitive data. Following are the approaches:

a) K-anonymity

K-anonymity [26, 28] and its alternative are widely used. The idea of *k*-anonymity is to modify the values of quasi-identifiers in the original data table; so that every tuple in the anonymized table is in distinguishable from at least *k*-1 other tuples along the quasi-identifiers. The anonymized table is called a *k*-anonymous.

DOB	Gender	Pin	Health Issue
02/02/1993	Female	226025	chest pain, short breath
23/12/1999	Male	226015	Flu
04/05/1989	Male	226025	Thyroid
03/11/1996	Male	200101	Hypertension
15/01/1994	Female	226027	Gastritis
30/02/1990	Female	226015	Hiv
31/08/1987	Male	226027	Cancer

(a) Actual Table

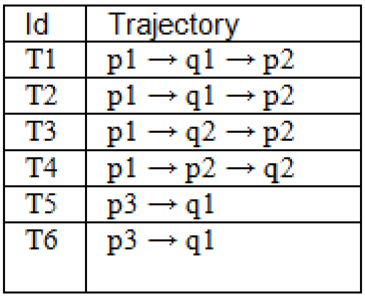
DOB	Gender	Pin	Health Issue
1993	Patient	2*****	chest pain, short breath
1999	Patient	2*****	Flu
1989	Patient	2*****	Thyroid
1996	Patient	2*****	Hypertension
1994	Patient	2*****	Gastritis
1990	Patient	2*****	Hiv
1987	Patient	2*****	Cancer

(b) Anonymous Table

(Example of 2-anonymity, where QI= {Date of Birth, Gender, Pin}, (a) Actual Table, (b) Anonymous table. Figure 2 provides an example of a table T that holds to k-anonymity. The quasi-identifieris QI= {DOB, Gender, and PIN}. Then, for every tuple contained in the table, the values of the tuple that comprise the quasi-identifier appear at least 2 occurrences of those records.

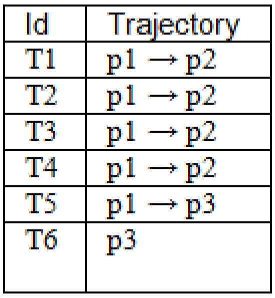
b) Trajectory Data

Today, with the growing interesting e- Government and e- commerce, sensitive data exchanged online, expanded availability of mobile devices with, location based services is becoming very popular in current years. For example, Social networking site Facebook where users ‘‘check-in’’ to different places like restaurants, clubs etc., to share their current position with friends. Other than check-ins number of users also shares their complete movement direction; these services are very trendy; their use may raise serious privacy concerns [30].Here authors concentrate on the privacy issues brought by releasing trajectory data of users. To offer location-based services, business entities (Telecommunication organization) and public entities (Transportation organization) gather a vast amount of user’s trajectory data, i.e. sequences of consecutive location readings along with time stamps. If the data collector releases such spatiotemporal data to an outsider sensitive information about users might be unveiled [26].



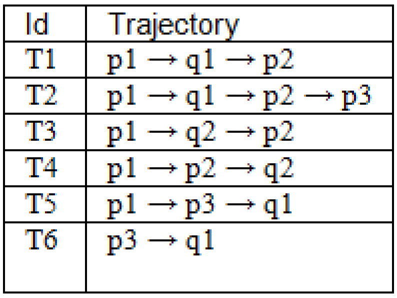
Id	Trajectory
T1	p1 → q1 → p2
T2	p1 → q1 → p2
T3	p1 → q2 → p2
T4	p1 → p2 → q2
T5	p3 → q1
T6	p3 → q1

(a) Actual Table



Id	Trajectory
T1	p1 → p2
T2	p1 → p2
T3	p1 → p2
T4	p1 → p2
T5	p1 → p3
T6	p3

(b) Attackers Knows



Id	Trajectory
T1	p1 → q1 → p2
T2	p1 → q1 → p2 → p3
T3	p1 → q2 → p2
T4	p1 → p2 → q2
T5	p1 → p3 → q1
T6	p3 → q1

(c) Transformed Table

If an attacker is familiar with his target A, repeatedly visited two places *p*1 and *p*3, then he can know for beyond any doubtthat the trajectory *T*3 corresponds to A since there is theonlytrajectory that goes through *p*1 and *p*3. While if some of the locations are suppressed, as shown in Fig. (a), Mark cannot differentiate between *T*3 and *T*4, thus the trajectory of A will not be disclosed [26].

c) Association Rule Mining

Association rule mining [26-27] is one of the most important data mining tasks, which aims at finding interesting associations and correlation relationships among large sets of data items. In general, the association rule mining process contains following two steps:

- Step 1: discover all frequent data set. The occurrence frequency of a data set is the number of transactions that contain the data set. A frequent dataset is a data set whose occurrence frequency is greater than a prearranged lowest support count.
- Step 2: strong association rules Generate from the frequent datasets that satisfy both a minimum confidence threshold and a minimum support threshold are called strong association rules.

Transaction ID	Actual Data	Modified Data
T1	P	PR
T2	PQR	PQ
T3	PQ	PQ
T4	PQR	PQR

d) Clustering

Cluster analysis is the technique of grouping a collection of data items into various groups or clusters so that items withina cluster have close similarity, but are very different to items in other clusters. Dissimilarities and similarity are based on the attribute values describing the items and often involve distance measures. Clustering methods can be classified into hierarchical methods, partitioning methods,density-based methods,etc.
Recent research on privacy-preserving clustering can be roughly classified into two categories, perturbation and secure multi-party computation (SMC) [26].

e) File Encryption

Encryption ensures confidentiality and privacy of user information. A hacker can steal confidential data which is present in the cluster. Hence, all the stored data should be encrypted. There are different encryption algorithms like DES, AES, etc. These algorithms use encryption keys to encrypt data. If an attacker accesses the encrypted data, then he cannot fetch meaningful information from it and misuse it. So the confidential data will be securely stored in an encrypted manner [8] [11].

f) Nodes Authentication

Authentication is a validating system or user’s identity at the time to accessing the system. Whenever a node enters in a cluster, it should be authenticated. In the case of a malicious node, it should not be permitted to join the cluster. Hadoop provides Kerberos as a primary authentication. Kerberos can be used to authenticate the authorized nodes from malicious ones [8, 11].

g) Honeypot Nodes

A honeypot is a unique security tool that you want the hacker to interact with it. In the cluster, Honeypot nodes should be present, which seem like an ordinary node but is a trap [11]. A honeypot works as an information system resource whose value relies on between the illegal or illicit use of that source [12].

h) Access Control

For a good security, in a heterogeneous environment, privacy and access control’s Integration will be compulsory. Data providers can control their sensitive data with the security policy. They will also control the mathematical bound on privacy violation that could take place. To prevent information leak, SELinux will be used. It is a Security-Enhanced Linux; in the Linux Kernel, it provides a feature for supporting access control security policy through the use of Linux Security Modules [11].

VII. CONCLUSION

A huge amount of data about individuals related to their medical, internet activity, social networking, energy usage, communication patterns and social interactions is called big data. It is created from different sources such as hospitals, social networking sites etc. From these sources, data is being collected and processed by various survey organizations, national statistical agencies, medical centers, The Web, and companies. Data has not a fixed source it is collected from various sources. To maintain big data in a heterogeneous environment is a difficult task.

The main focus of this paper is to describe security issues that are associated with big data in the context of user’s authentication, authorization and security measure techniques to overcome these issues in big data are discussed. Big data is widely used in industry and research aspects; it is not only about storing or retrieving data. Analysis of the same is required to make it useful and to extract the required information from it. Therefore security is an important aspect for organizations running on these environments. Using proposed approaches, big data can be secured for complex business operations

VIII. REFERENCES

[1] Bhosale S Harshawardhan, et.al. "A Review Paper on Big Data and Hadoop." International Journal of Scientific and Research Publications, Volume 4, Issue 10, 2014 October, PP. 1-7.
[2] Bello-Orgaz, et.al. "Social, big data: Recent achievements and new challenges." Information Fusion: pp. 45-59, the year 2015.
[3] Shirudkar Kalyani, et.al, “Big-Data Security”, International Journal of Advanced Research in Computer Science and Software Engineering, pp. 1100-1009, the year2015.

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

[4] Kshetri, Nir. "Big data′ s impact on privacy, security and consumer welfare". Telecommunications Policy 38.11, pp. 1134-1145, year 2014.

[5] Gandomi, Amir, and Murtaza Haider. "Beyond the hype: Big data concepts, methods, and analytics," International Journal of Information Management, 2015, pp.137-144.

[6] Alexandru Adrian TOLE, "Big Data Challenges", Database Systems Journal vol. 4, 2013, pp. 31-40.

[7] UN Statistical Commission. "Big data and modernization of statistical systems–Report of the Secretary General." UN Economic Social Council, 2014 March.

[8] Sharma, Priya et.al, "Securing Big Data Hadoop: A Review of Security Issues, Threats, and Solution", IJCSIT) InternationalJournal of Computer Science and Information Technologies 5.2, 2014.

[9] Nedelcu, Bogdan, "About Big Data and its Challenges and Benefits in Manufacturing", Database Systems Journal 4.3, 2013, pp. 10-19.

[10] Savant, et.al, "Approaches to Solve Big Data Security Issues and Comparative Study of Cryptographic Algorithms for Data Encryption", International Journal of Engineering Research and General Science Volume 3, 2015, pp. 425-428.

[11] In Kullu, et.al, "Security Issues Associated with Big Data in Cloud Computing."International Journal of Network Security & Its Applications (IJNSA), 2014.

[12] Spitzner, L, "Honeypots: Catching the insider threat", Computer Security Applications Conference, Proceedings. December IEEE, 2013, pp. 170-179.

[13] Kaisler, Stephen, et al. "Big data: Issues and challenges moving forward."System Sciences (HICSS), 2013 46th Hawaii International Conference on. IEEE, 2013 pp. 995-1004.

[14] Amir Gandomiet.al, "Beyond the hype: Big data concepts, methods, and analytics" International Journal of Information Management, 2013 pp. 137–144.

[15] Dumbill, Edd. "What is big data? An Introduction to the Big Data Landscape". Strata, 2012..

[16] Gai, Keke, et al. "Proactive attribute-based secure data schema for mobile cloud in financial industry." High Performance Computing and Communications (HPCC), 2015 IEEE 7th International Symposium on Cyberspace Safety and Security (CSS), 2015 IEEE 12th International Conference on Embedded Software and Systems (ICESS), 2015 IEEE 17th International Conference on. IEEE, 2015.

[17] Demchenko Yuri et.al, "Defining architecture components of the Big Data Ecosystem", In Collaboration Technologies and Systems (CTS), 2014 International Conference on IEEE, 2014 pp. 104

[18] Hashem, et al. "The rise of "big data" on cloud computing: Review and open research issues." Information Systems 47, 2015 pp. 98-115.

[19] Kaisler, et al. "Big data: issues and challenges moving forward." System Sciences (HICSS), 2013 46th Hawaii International Conference on. IEEE, 2013, pp. 995-1004.

[20] Miloslavskaya Natalia, et al. "Big Data information security maintenance." Proceedings of the 7th International Conference on Security of Information and Networks. ACM, 2014.

[21] Colin Tankard, "Big data security", Network Security, 2012, pp. 5-8.

[22] Achana RA, et.al, "A novel data security framework using E-MOD for big data". IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE), 2015 Dec, pp. 546-551.

[23] OzkoseHakan et.al, "Yesterday, Today and Tomorrow of Big Data", procedia- social and Behavioral Sciences, 2015, pp. 1042-1050.

[24] KatalAvita, et.al, "Big Data: Issues, Challenges, Tools and Good Practices", Contemporary Computing (IC3), 2013 Sixth International Conference on. IEEE, 2013, pp. 404-409.

[25] Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Khan, S. U. (2015). The rise of "big data" on cloud computing: Review and open research issues. Information Systems, 47, 98-115.

[26] Xu, L., Jiang, C., Wang, J., Yuan, J., & Ren, Y. (2014). Information security in big data: privacy and data mining. IEEE Access, 2, 1149-1176.

[27] Rathna, S. Selva, and T. Karthikeyan. "Survey on recent algorithms for privacy preserving data mining." International Journal of Computer Science and Information Technologies 6.2 (2015): 1835-40.

[28] Wernke, M., Skvortsov, P., Dürr, F., & Rothermel, K. (2014). "A classification of location privacy attacks and approaches. Personal and Ubiquitous Computing, 18(1), 163-175.

[29] DinushaVatsalan, Peter Christen , Christine O'Keefe‡, and Vassilios S. Verykios, "An Evaluation Framework for Privacy-Preserving Record Linkage", Journal of Privacy and Confidentiality, 2014: 35-75.

[30] Medforth, Nigel, and Ke Wang. "Privacy risk in graph stream publishing for social network data." Data Mining (ICDM), 2011 IEEE 11th International Conference on. IEEE, 2011.

[31] Alshboul, Yazan, Wang Yong, Nepali Kumar Raj, "Big Data Life Cycle: Threats and Security Model.", Americas Conference on Information Systems, the year 2015, pp 1-7.

[32] Miloslavskaya Natalia, et al. "Big Data information security maintenance."Proceedings of the 7th International Conference on Security of Information and Networks. ACM, 2014

[33] http://www.computerhope.com/jargon/u/unauacce.htm

[34] http://www01.ibm.com/software/data/infosphere/hadoop/whatis-big-data-analytics.html.

CITE AN ARTICLE

K., Agrawal, A., & Khan, R. A. (2017). AN OVERVIEW OF PRIVACY PRESERVATION IN BIG DATA. *INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY*, *6*(7), 61-71. doi:10.5281/zenodo.822970