Big Data Security and Privacy Challenges

[1]Abdullah Al-Shomrani , [2]Fathy Eassa, [3]Kamal Jambi
Computer Science
King AbdulAziz University, Jeddah, Saudi Arabia

***Abstract*— Big Data security and privacy is a big challenge for both data owner and service providers. Big Data has become a necessity for business, researchers, healthcare, and government agencies. However, the tools and technologies that are being developed to manage this volume of data are not designed to address security and privacy requirements. In this research we will define the sources of the Big Data and the characteristics and finally what are the security and challenges facing Big Data.**

***Index Terms*— Big Data, Privacy, Security, Chrematistics.**

I. INTRODUCTION

Big Data is magnified by velocity, volume and variety [1]. The volume of data is increasing every second from different input resources. Big Data is the terminology used to describe huge volumes of data that are very large and cannot be processed using normal databases and software [2]. The tremendous growth in data size due to the velocity of data collection and processing of data inputs coming from the widespread deployment of connected devices such as automobiles, smart phones, RFID readers, web cameras, and sensor networks requires huge resources to handle this size of data with ensuring the security of data.

Devices such as the above mentioned continuously generate data streams without human intervention. These unstructured data sources contribute to a much higher variety of data types. The massive volume, velocity, and variety of data have an enormous impact to existing security solutions that were not designed and built with Big Data in mind. Over the past two years, 90% of the world's data has been generated [3].

The speedy rise of digital technology, such as digital sensors, smart devices, networks, digital logs and more have made the creation of data, its storage, use and protection the more important aspect of society and business enterprises over the last few years. Over the long term, the extensive use of these devices and technologies have led to the generation of exponential volumes of information, which are of varying nature, sizes, forms, and complexities. Big Data arises because of the integration of the sorts of data with highly diversified characteristics and features, and the rate at which data and complexity are generated requires storage and handling that is beyond the means of conventional techniques and procedures [4].

The amount of data has doubled every two years and requires different storage strategies with different storage media [5]. Big data security and privacy are big challenges for customers and service providers [1]. Eighty percent of large organizations will endure major security issues with big data by 2016[6]. Most of these data are not in standard forms, which make it more difficult to analyze with the available tools of today [7]. Big Data, one of the most recent technology trends in 2013, is facing big challenges with security and privacy which is threatening to slow this momentum.

The Internet of Things (IoT) will be connecting 26 billion devices by 2020, generating huge amounts of data by using devices with real-time monitoring which is a big challenge for protecting client privacy [8]. Personal information and client data related to business, combined with social information, will allow the market and unauthorized access to this information [9]. Every modern electronic device in the world today is connected to the internet and is collecting, generating and storing data which is huge.

II. BIG DATA SOURCES

Large data sources can be classified as follows: [10]

- Administrative (arising from the administration of a program, be it governmental or not), e.g. electronic medical records, hospital visits, insurance records, bank records, food banks, etc.
- Commercial or transactional: (arising from the transaction between two entities), e.g. credit card transactions, on-line transactions (including from mobile devices), etc.
- Sensors, e.g. satellite imaging, road sensors, climate sensors, etc.
- Tracking devices, e.g. tracking data from mobile telephones, GPS, etc.
- Behavioural, e.g. online searches (about a product, a service or any other type of information), online page view, etc
- Opinion, e.g. comments on social media, etc.
- As we can see in Figure 1 which summarize the sources of big data.

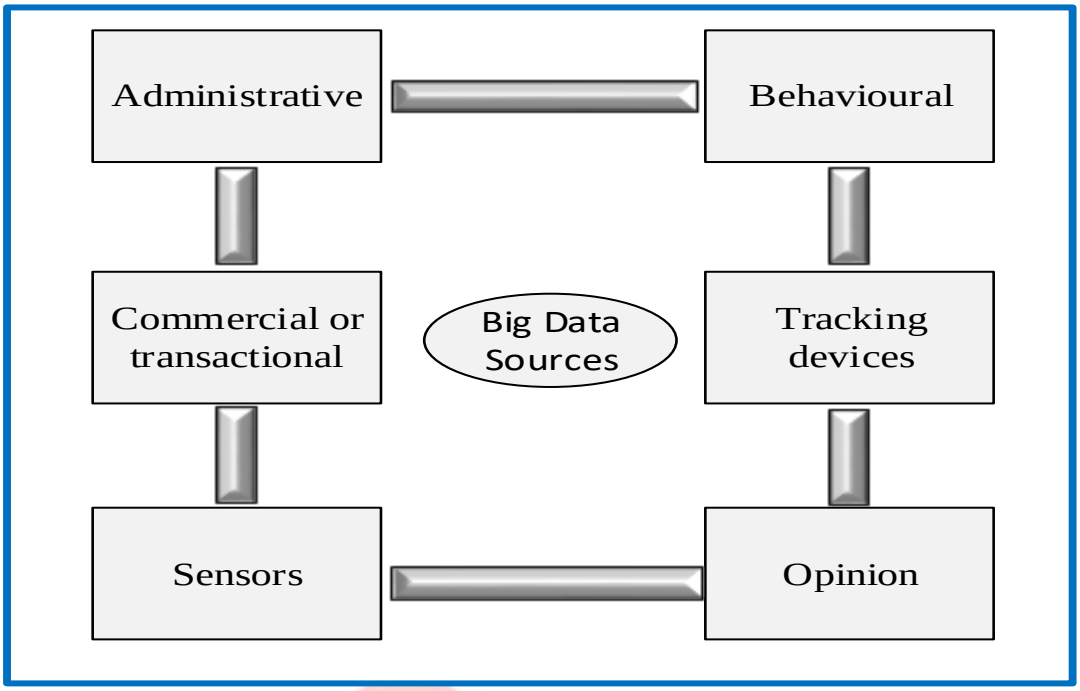

Figure1: Big Data Sources

III. CHARACTERISTICS OF BIG DATA

Big data is a term used to describe the collection of large and complex data sets that are difficult to process using on hand database management tools or traditional data processing applications. Big data spans across seven dimensions which include volume, variety, volume, value, veracity, volatility and complexity [11] as can be seen in Figure 2.

- Volume: The volume of data here is very huge and is generated from many different devices. The size of the data is usually in terabytes and petabytes. It refers to the size of data being created from all the sources including text, audio, video, social networking, research studies, medical data, space images, crime reports, weather forecasting and natural disasters, etc.
- Velocity: This describes the real-time attribute found in some of the data sets for example streaming data. The result that misses the appropriate time is usually of little value.

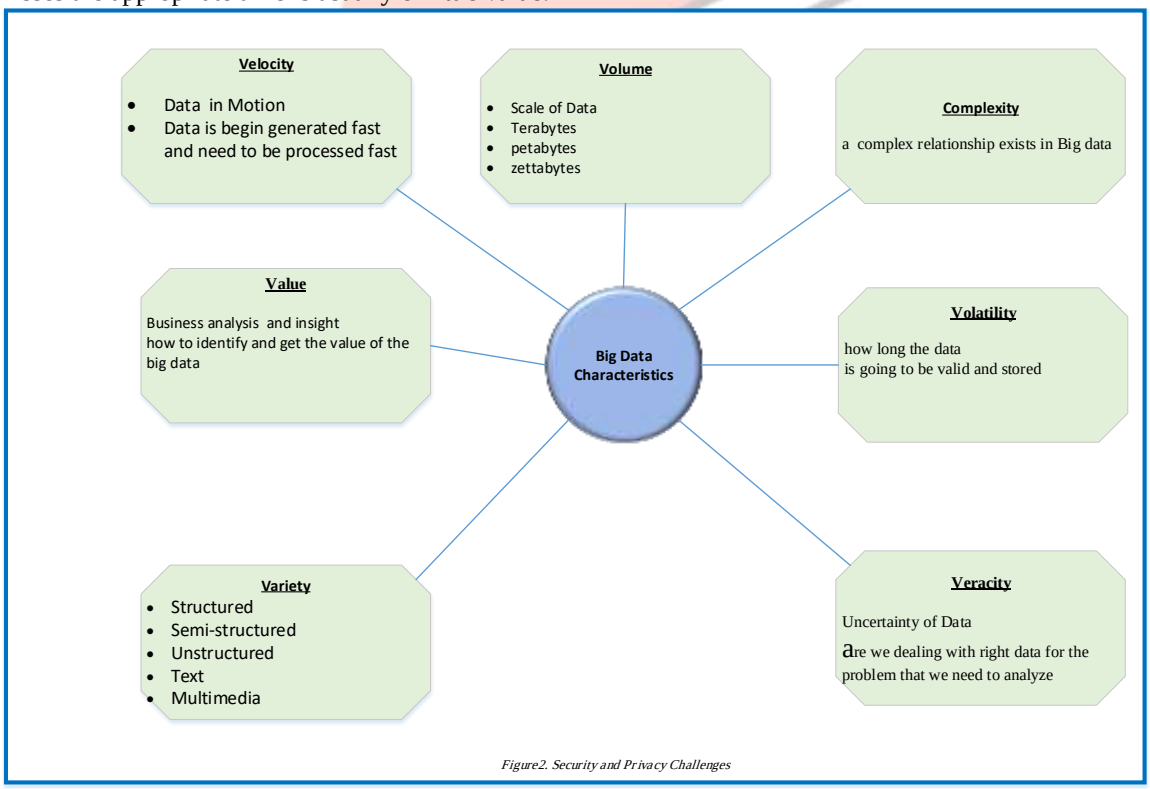

Figure2. Security and Privacy Challenges

- Variety: Big data consists of a variety of different types of data i.e. structured, unstructured and semi-structured data. The data maybe in the form of blogs, videos, pictures, audio files, location information etc.

- Value: This refers to the complex, advanced, predictive, business analysis and insights associated with the large data sets.
- Veracity: This deals with uncertain or imprecise data. It refers to the noise, biases and abnormality in data. This is where we find out if the data that is being stored and mined is meaningful to the problem being analyzed.
- Volatility: Big Data volatility refers to how long the data is going to be valid and how long it should be stored.
- Complexity: A complex dynamic relationship often exists in Big data. The change of one data might result in the change of more than one set of data triggering a rippling effect

IV. DATA SECURITY AND PRIVACY CHALLENGES

The Top 10 security and privacy challenges to overcome in Big Data according to survey conducted by Cloud Security Alliance members and security practitioner-oriented trade journals the list of high-priority security and privacy challenges [38]. The following are the top ten security and privacy challenges.

- Secure computations in distributed programming frameworks
- Security best practices for non-relational data stores
- Secure data storage and transactions logs
- End-point input validation/filtering
- Real-time security monitoring
- Scalable privacy-preserving data mining and analytics
- Cryptographically enforced data centric security
- Granular access control
- Granular audits
- Data provenance

The Cloud Security Alliance grouped the above challenges into four broad components as showed in figure 3.

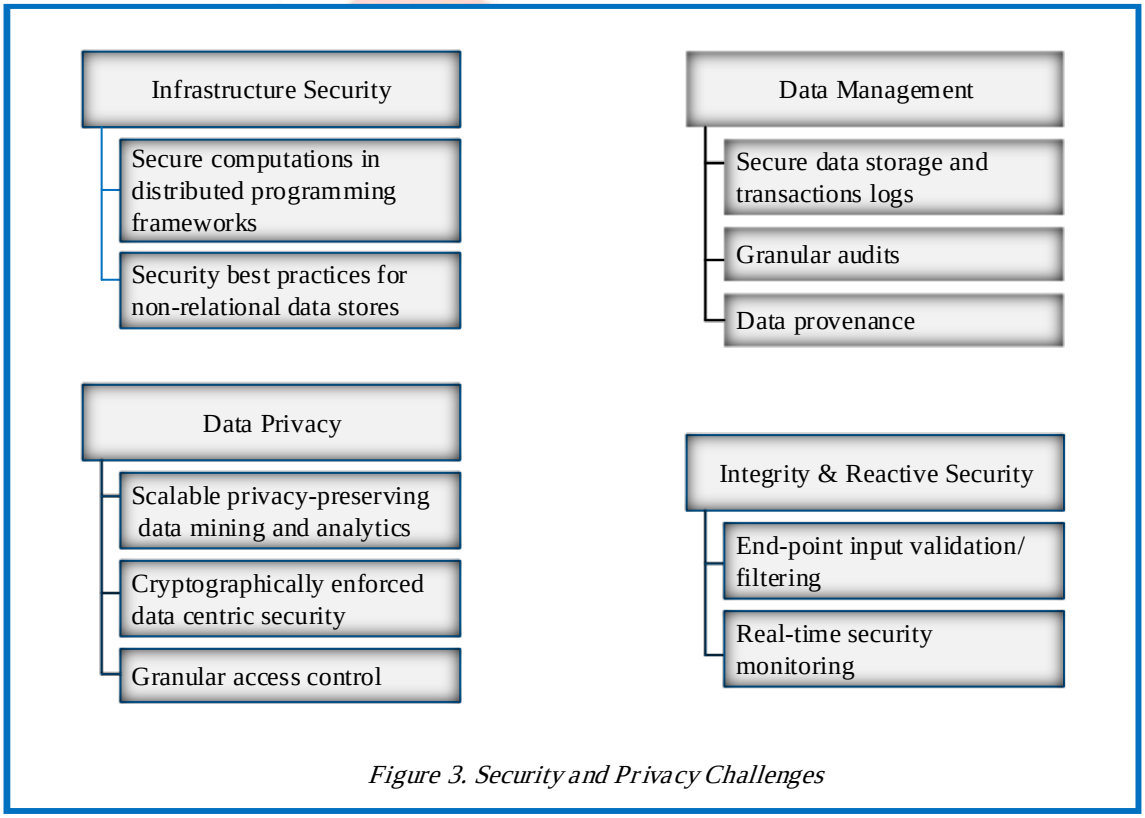

Figure 3. Security and Privacy Challenges

There are many challenges facing Big Data; security and privacy are just some of those challenges. It has been reported that the growth of big data increases the threats to the existing security of the information. Big data privacy is very important for the data owner and for the service provider [12].

Big data sizes are continuously increasing from terabytes in 2012 to nearly 44 zettabytes by the year 2020 in a single data set. In [13] they categorized the security challenges as follows: Infrastructure Security, Data Privacy, Data Management and Integrity and Reactive Security.

Protection of sensitive and private big data is needed to protect the privacy of users. The protection of sensitive data requires the ability to be handled from different angles of security: privacy, authenticity, integrity and access control are the critical security goals for big data [14]. Security of data required to have the system satisfy the following: confidentiality, integrity, and availability [15].

There are certain requirements for security of big data in different areas of application that include government, social networks, healthcare and other applications. Protecting Big Data while it is in storage has been a challenge to most organizations due to its size and volume, especially as they attempt to maximize data efficiency and performance. The optimal data protection is to

ensure that when it falls into unauthorized hands, it is meaningless. Encryption is a method that can provide that protection. However, securing data storage need to in place prior of distributing the data across the cloud [16].

Personal and private information are collected and analyzed by unauthorized entities to gain knowledge and make predictive analysis that will direct people without their awareness [17].

There are two sources of security threats for data storage. The first threat comes from the service provider. The second source is an adversary that is economically motivated and has the capability to compromise the data storage servers [1]. Lack of secure communication between nodes is a threat for big data during transmission that can be solved by using a secure link such as SSL/TLS. Data poisoning is another threat which means an adversary may compromise data in transmission. Insurance of the source of data is a major challenge for data security to detect fraud.

Data privacy [18] is a comprehensive process integrating protection of data at the data generation, storage and processing stages throughout the big data lifecycle. As a result, privacy requirements for data directly correlates with the unique requirements of each stage. In modern society, businesses value privacy and invest a lot of resources in the interest of protecting their clients. Similarly, there is an increase in the use of cryptographic systems to complement the human efforts in fostering data's physical security. A shift toward privacy preservation has gained considerable traction in recent years by incorporating privacy throughout the big data lifecycle. Moreover, big data security fosters data privacy through implementing a variety of encryption techniques. The approaches include Identity and attribute based encryptions and storage path encryption systems.

Encryption of all data is expensive and overkill as it assumes all data as sensitive. As an example, if we consider the data for a patient, the personal details are sensitive while the illness is not. This means we need to secure the association between the identity of the patient and the illnesses. Encryption of data depends on the sensitivity, based on the user privacy rules, therefore data is separated and encrypted when it is necessary

Big data security has grown fast as a significant concern for clients over the past few years. 88% of the clients were substantially worried on their data privacy [1]. The security issues are increasing and it is happening because of the increasing usage of big data through adaptation of this technology. There are many benefits of big data. Although, it is vulnerable to attacks. Attackers are consistently trying to find loopholes to attack the big data storage.

Current technology was not built for big data security. Data is stored in plain text; wherein critical information can be easily stolen by hackers. Logging to critical data is not logged in which means any abused of data cannot be identified. This technology is used by most businesses to store and analyze their own data and their customer's data which makes privacy and security very important in gaining the confidence of customers. Hence, there is a need for investing, studying, and understanding the challenges and providing better solutions to secure big data.

There are four safety factors to secure sensitive data [19] which are:

- Security issues when transmitting sensitive data from a data owner's local server to a big data platform.
- Securing sensitive data while computing and storage
- Sensitive data security issues on the cloud platform.
- Sensitive data destruction.

West in [20] defined privacy as "Privacy is the claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated to others."

Big data security and privacy are big challenges for customers and service providers [21]. Eighty percent of large organization will endure suffering from major security issues with big data by 2016[22]. Most of these are not in standard forms which make it more difficult to analyze with the available tools of today [23]. Ensuring privacy and isolation of data and resources by using adequate mechanisms for authentication and authorization [24] [25]. Beneficial information can be extracted and analyzed from digital information collected from different sources such as credit card companies, government agencies, banking, and health care. Regardless of the usage of the information and the benefit, privacy and security is the main challenge in big data. The privacy problem is main concern these days that comes because of the huge amount of personal information available on internet as digital information.

Nisbet, R., et al. [26] Asserts that data leakage creates a primary risk in the implementation of big data technologies by creating a vulnerability for more serious attacks. Primarily, data leakage occurs inadvertently and is attributed to flaws in the design of big data systems. Insecure web applications may process sensitive information and leading to unintended data leakage. Consequently, the data is stored in an insecure location and can be exploited by malicious attackers or user applications. This niche in big data security is created by the poor prioritization of data security in design and subsequently the limited compliance to the industry security protocols and standards of practice in web application development

In [27] has interpreted that data privacy challenges entail the protection of personally identifying information to the best extents possible while at the same time permitting the performance of vital analysis on the same. This includes ensuring scalability as well as composable privacy, and ensuring data mining and analytics are jot undesirably exposed or disclosed to parties that do not have any urgent and justifiable need of such access. The enforcement of data privacy also calls into action the use of cryptographic encryption of data to take meaning out of it without success to the relevant keys needed to decrypt such data. Granular access has also been proposed as a possible robust measure in the protection of such privacy.

V. ACCESS CONTROL

Access control is one of most for challenges big data [28]. Security policy allows the organization to safe guard their data. Authorization in Big Data is more complicated because it considers the content [29].

Authentication

Authentication is the process of validating a user's physical claimed identity or the digital identity of a process or a computer. User authentication can be categorized into three main categories: authentication by knowledge - i.e., what the user knows such as a password (memo metrics), authentication by possession - i.e., what the user has, such as a smartcard token (econometrics), and authentication by characteristics - i.e., biometrics such as fingerprints, retinal, iris, voice, face, handwritten (biometrics)[30] . These authentication approaches can be combined or used separately, depending on the demanded level of functionality and security. Biometrics authentication offers authentication based on the measurement of unique physiological characteristics of a user, such as fingerprints and face recognition [31]. Authentication in Big Data is complicated as data created from multiple sources authenticating itself to a common server [32].

Authorization

Authorization is the process of granting or denying access to the platform resources based on the identity of users. An authorization module enforces security policies that are configured for each role in the active security domain where authentication performed. The authorization process checks permission rights when an authenticated user requests access to a service.

The authorization system retrieves the groups' information through the custom authentication realm for users with the valid authenticated sessions. For example, when a user is authenticated, a permission check retrieves all the user's related groups. If the requested action is permitted on a service or a resource, the user will be granted access.

VI. FRAGMENTATION

Fragmentation is a technique to partition a given relation R or set of data into two or more Partitions such that the combination of the partitions provides the original database without any loss of information.

Given a relation schema R, a set C of well defined policy constraints, and a set $A \subseteq R$ of attributes to be fragmented, a fragmentation of R on A is a set of fragments F={F1,...,Fr}.

1. Relational data fragmentation:

There are three fragmentation types in the relational data [33]: vertical fragmentation, horizontal fragmentation and hybrid fragmentation. Verticals fragmentation (VF) will be used for the relational data. Vertical fragmentation allows a relation to be partitioned into disjoint sets of columns or attributes. Each partition must include a key attribute(s) of the table. VF provides more security to the data than the horizontal fragmentation (HF). HF allows a relation to be partitioned into disjoint tuples or instances. The issue with HF that each partition can provides useful information by itself. For our framework VF will be used which give more security to sensitive data. The objective of vertical fragmentation is to partition a relation into a policy constraint will be part of the data. When data received, the algorithm will scan the security set of smaller relations so that each part will not provide knowledge without the other parts. The security constraints and bullied the association between the data and the security rules. The data will be passed to the fragmentation algorithms. The same process will be done for the data that have been saved without applying the security approach. The framework will scan the data and apply the fragmentation and relocate the data with encryption for the sensitive data.

- There are two kinds of confidentiality requirements that can be applied to the data, attribute is sensitive or the association among some attributes is sensitive [34].
- Sensitive attributes. Some attributes are sensitive and their values should be maintained confidential. Simple examples of such attributes are SSN, credit card numbers, emails or telephone numbers and similar attributes whose values should not be released.
- Sensitive associations. In some cases, what is sensitive is the association among attributes values rather than the values of an attribute. For instance, the names of patients in a hospital may be considered not sensitive, and so the diseases treated by the hospital; however, the specific association between individual patients and their illnesses is sensitive and should be maintained confidential.

2. Text data fragmentation:

There will be two different approaches, the first one, sensitive part of the data will identify by the user using indication of the start and the end of the sensitive part of the text as showed in example 1. The second approach, in case of the sensitive data was not pointed by the user; the algorithm will scan through the text and identify the key sensitive word and mark the text based on the security constraints requested by clients. Text segmentation is the process of partition text into segments, such as tokens, phrases, or topics [35] [36].

3. XML fragmentation:

XML will be used as one of the semi structure data type. it split an XML document into a new set of XML documents. Their main objective is either to improve XML query performance [37] or to distribute or exchange XML data over a network. This fragmentation splits XML document elements and assigns a reference to each sub-element.

In this type of data fragmentation algorithm will use the concept of hole and filler. Data sensitive tag will identify if the data is sensitive or not. When the data is sensitive the sensitive data will be replaced by hole id. Sensitive data will be extracted and given id as filler. The extracted data will be fragmented and then encrypted. For the data that the sensitive data is not identified the algorithm will scan through the XML file and identify the sensitive data based on the user policy constraints.

REFERENCES

[1] Cloud Security Alliance "Top ten big data security and privacy challenges" November 2012. https://www.isaca.org/Groups/Professional-English/big- data/GroupDocuments/Big_Data_Top_Ten_v1.pdf

[2] Venkata Narasimha Inukollu, Sailaja Arsi ,and Srinivasa Rao Ravuri, "Security issues associated with big data in cloud computing", International Journal of Network Security & Its Applications (IJNSA), Vol.6, No.3, May 2014 .

[3] Kudakwashe Zvarevashe1, Mainford Mutandavari2, Trust Gotora3; "A Survey of the Security Use Cases in Big Data"; International Journal of Innovative Research in Computer and Communication Engineering; Vol. 2, Issue 5, May 2014.

[4] Eweka R. Osawaru, Riyaz Ahamed A. H. "A Highlight of Security Challenges in Big Data", International Journal of Information Systems and Engineering (online), Volume 2, Issue 1 (April 2014).

[5] Klaus Engelhardt " Secure data storage an overview of storage technology", white paper storage technology 2008.

[6] Chris Marrison "Gartner warns of big data security problems", Network-Security,2014.

[7] Elmustafa Sayed Ali Ahmed1 and Rashid A.Saeed, "A Survey of Big Data Cloud Computing Security," International Journal of Computer Science and Software Engineering (IJCSSE), Volume 3, Issue 1, December 2014.

[8] K. Harsh and S. Ravi, "Big Data Security and Privacy Issues in Healthcare", 2014, IEEE International Congress on Big Data, pp. 762-765, (2014).

[9] Abeer M. AlMutairi, Rawan Abdullah, Jayaprakash Kar; "Security and Privacy of Big Data in Various Applications"; International Journal of Big Data Security Intelligence Vol. 2, No. 1 (2015).

[10] united nations economic commission for europe; conference of european staticians; what does big data mean for officila statistics; 10 march;2013.

[11] Xiaoxue Zhang, Feng Xu, "Survey of Research on Big Data Storage", 2013 12th International Symposium on Distributed Computing and Applications to Business, Engineering & Science

[12] J. Surana, A Khandelwal, A. Kothari " Big Data Privacy Methods" international journal of engineering devlopment and reserch, 2017 IJEDR | Volume 5, Issue 2 | ISSN: 2321-9939.

[13] Puneet Goswami, " A Survey on Big Data & Privacy Preserving Publishing Techniques " , Advances in Computational Sciences and Technology ISSN 0973-6107 Volume 10, Number 3 (2017) pp. 395-408

[14] 11. A, Katal, Wazid M, and Goudar R.H. "Big data: Issues, challenges, tools and Good practices."; Noida: 2013, pp. 404 – 409, 8-10 Aug. 2013.

[15] Atallah, M.J., Frikken, K.B., Goodrich, M.l.T., Tamassia, R.: Secure biometric authentication for weak computational devices. In: Patrick, A.S., Yung, M. (eds.) FC 2005. LNCS, vol. 3570, pp. 357–371. Springer, Heidelberg (2005)Google Scholar

[16] Liu, H., Wang, H., Chen, Y., "Ensuring Data Storage Security against Frequency-Based Attacks in Wireless Networks", In: Rajaraman, R., Moscibroda, T., Dunkels, A., Scaglione, A. (eds.) DCOSS 2010. LNCS, vol. 6131, pp. 201–215. Springer, Heidelberg (2010).

[17] Neha Upadhyay , Ajay Kumar; "A Framework based on Authentication and Authorization to ensure Secure Data Storage in Cloud"; International Journal of Computer Applications (0975 – 8887), Volume 90 – No 15, March 2014.

[18] Priyank Jain, Manasi Gyanchandani and Nilay Khare., Big data privacy: a technological perspective and review, Journal of Big Data, Springer 26 November 2016.

[19] Xinhua Dong, Ruixuan Li , Heng He, Wanwan Zhou, Zhengyuan Xue, and Hao Wu, "Secure Sensitive Data Sharing on a Big Data Platform ", TSINGHUA SCIENCE AND TECHNOLOGY I S SNll1 0 0 7 - 0 2 1 4ll0 8/ 1 1llp p 7 2- 8 0 Volume 20, Number 1, February 2015

[20] West A. Westin, "Privacy and Freedom", Washington and Lee Law Review, Volume 25, Issue 1.

[21] Zeng G. (2015). Big data and information security, international journal of computational engineering research, vol. 5 Issue. 6 pp. 17-21

[22] Chris Marrison "Gartner warns of big data security problems", Network-Security,2014.

[23] Elmustafa Sayed Ali Ahmed1 and Rashid A.Saeed, "A Survey of Big Data Cloud Computing Security," International Journal of Computer Science and Software Engineering (IJCSSE), Volume 3, Issue 1, December 2014.

[24] H. Li, Y. Dai, L. Tian, and H. Yang, "Identity-based authentication for cloud computing," Cloud Computing, pp. 157–166, 2009.

[25] W. Jansen, "Cloud hooks: Security and privacy issues in cloud computing," in System Sciences (HICSS), 2011 44th Hawaii Int. Conf. on. IEEE, 2011, pp. 1–10.

[26] Nisbet, R., Elder, J. and Miner, G. (2009) Handbook of Statistical Analysis and Data Mining Applications. Academic Press.

[27] Ma L., Pei Q., Leng H., & Li H. (2015). Survey of security issues in big data, Radio communications technology, vol. 41 Issue 1 pp. 1-7

[28] V. Hu and K. Scarfone, "Guidelines for Access Control System Evaluation Metrics," NIST Interagency Report 7874, Gaithersburg, MD, USA, 2012.

[29] An Access Control Scheme for Big Data Processing Vincent C. Hu, Tim Grance, David F. Ferraiolo, D. Rick Kuhn National Institute of Standards and Technology Gaithersburg, MD, USA vhu, grance, dferraiolo, kuhn@nist.gov

[30] L. Cranor and S. Garfinkel, Security and Usability: Designing Secure Systems
that People Can Use. O'Reilly Media, 2005.

[31] M. J. Atallah, K. B. Frikken, M. T. Goodrich, and R. Tamassia, "Secure biometric authentication for weak computational devices," in *Proceedings of the* 9th International Conference on Financial Cryptography and Data Security, FC'05, (Berlin, Heidelberg), pp. 357–371, Springer-Verlag, 2005.

[32] M Jhaveri, D Jhaveri, N Shekokar, "Big Data Authentication and Authorization using SRP Protocol " , International Journal of Computer Applications (0975 – 8887) Volume 130 – No.1, November 2015
[33] A. Koreichi and B. L. Cun. On data fragmentation and allocation in distributed object oriented databases. Technical report, PRiSM, Versailles University, France, 1997.
[34] S. De Capitani, di Vimercati, R. Erbacher, S. Foresti, S. Jajodia, G. Livraga, P. Samarati, A. Aldini, J. Lopez, F. Martinelli, "Encryption and fragmentation for data confidentiality in the cloud" in Foundations of Security Analysis and Design VII ser. Lecture Notes, Computer Science, Springer International Publishing, vol. 8604, pp. 212-243, 2014.
[35] Y. Zhang, S. Vogel, and A. Waibel. Integrated phrase segmentation and alignment algorithm for statistical machine translation. In Proceeding of International Conference on Natural Language Processing and Knowledge Engi neering (NLP-KE), 2003.
[36] G. Salton, J. Allan, and A. Singhal. Automatic text decomposition and structuring. Information Processing and Management: an International Jour- nal, 32(2):127{138, 1996.
[37] A. Bonifati and A. Cuzzocrea. Efficient Fragmentation of Large XML Documents. In 18th International Conference on Database and Expert Systems Applications (DEXA 07), Regensburg, Germany,volume 4653 of LNCS, pages 539–550. Springer, 2007.
[38] Cloud Security Alliance "Top ten big data security and privacy challenges" November 2012. https://www.isaca.org/Groups/Professional-English/big- data/GroupDocuments/Big_Data_Top_Ten_v1.pdf.