ICIC Express Letters
Volume **12**, Number **9**, September **2018**
 ISSN 1881-803X
pp. **947–954**

SECURITY CHALLENGES AND PRECAUTIONARY MEASURES: BIG DATA PERSPECTIVE

Gayatri Kapil, Alka Agrawal and Raees A. Khan

Department of Information and Technology
Babasaheb Bhimrao Ambedkar University (A Central University)
Vidya Vihar, Raebareli Road, Lucknow 226025, India
gayatri1258@gmail.com; alka_csjmu@yahoo.co.in; khanraees@yahoo.com

Received January 2018; accepted April 2018

Abstract. *Big data is a combination of different datasets which is growing tremendously due to the digital environment created by various sources including mobile devices, servers, digital data of population collected by the government, call centres, etc. It is also important to recognize that most of this evolution is the outcome of an explosion in the number of devices located at the periphery of networks including embedded sensors, smartphones and personal computers, etc. This paper revolves around the big data analytics, its importance & applications. It also enumerated the directions to be taken while using big data along with security measures. In addition, the paper proposed an approach which can enhance the security and privacy of big data.*
Keywords: Big data dimensions, Big data analytics, Hadoop and security issues

1. **Introduction.** Big data is one of the most happening research areas. It is a huge amount of data which comes from various sources in the form of logs, blogs, emails, and other information [1]. It is in the form of structured, unstructured, and massive homogenous & heterogeneous [3]. Nearly five years ago, personal computer storage was tens to hundreds of gigabytes. In 2011, IDC's Digital Universe Study predicted that between 2009 and 2020 digital information data would grow by 44% from 0.8 ZB to 35 ZB [2]. Due to its nature, a better and modified model to handle and transfer the big data over the network is required [3]. Experts suggest that big data mining will provide the ability to take out the useful information from large datasets due to its characteristics like volume, variability and velocity. The same was not possible before [9]. As big data is different from other data in terms of volume, velocity, variety, and value [7], its processing and moreover security & privacy protection also become difficult for the government and the entrepreneurs. Because of its huge volume, traditional methods from managing to extract and then analysis are not very useful as they may not provide the accurate result for decision making, etc. Therefore, a new and capable processor is to be required which overcomes the various problems arisen during the processing of big data by traditional techniques. Apache Hadoop is an open source framework that can solve the big data issues like processing and managing within a tolerable time limit along with extensive analysis. The privacy protection of data is increased in the context of big data [9,11]. Privacy & security management is a major problem in technical and social networks. Thus, big data security is a moving target and requires more attention and focus. Big data need security & privacy protection from unauthorized modifications. In this paper, initially we have discussed about the big data analytics and its importance & uses and enumerated the directions to be taken while using the big data including security & privacy measures. At the end, the paper proposed an approach which can enhance the security and privacy of big data.

DOI: 10.24507/icicel.12.09.947

2. **Importance and Applications of Big Data Technology.** The time is not so far when big data will become a key basis for competition and growth of individual firms. In the view of growing competition and the ability to capture the relevant information, all companies have to take big data seriously [6]. This initiative will also lay the groundwork for complementary big data activities, such as big data infrastructure projects, platforms development and techniques in settling complex and data-driven problems in sciences & engineering. This is very helpful for the exploitation of information aspects and the creation of new facets for facilitator of decision-makers. Recently, many US government agencies, such as the National Institute of Health (NIH) and the National Science Foundation (NSF) have ensured that the use of data-intensive decision of big data has a profound effect on their future development. Consequently, they are trying to develop large data technologies and techniques to ease their mission after passing the big data initiative by the US Government on a large-scale [6]. According to the McKinsey Institute report [7], the effective use of big data has implicit benefits to transform economies and give a new wave of productive development. The advantages of valuable knowledge beyond the big data will be the basic competition for today's enterprises and will create new competitors, who are capable of attracting important skills employees on big data. Researchers, policymakers and decision makers have to recognize the ability to use big data to highlight the next wave of development in their areas. There are many benefits in the business segment, which can be obtained by big data, to increase operational efficiency, inform strategically, develop better customer service, find and develop new products & services, new customers and markets identity, etc. are included. Some applications of big data are explained below.

2.1. **Understanding the customers.** Every customer has a big problem which is very difficult to solve. They look at many things before buying, post about their purchases on social media and they want earnest thanks for buying a product. Big data allows you to profile these extremely vocal and playful small 'jingles' in a far-reaching way so that one-on-one, real-time interaction can be made with them. This is not really a luxury, if they are not treated as they want, they will leave in a very short time. Take an example of a customer who enters a bank, big data tool allows the banking staff to check his/her profile in real time to know which relevant products or services they can advise them. Big data will also have an important role in digital and physical shopping areas: a retailer can recommend a proposal to a mobile company, which indicates a certain demand of consumers in the social media [15].

2.2. **Interest based recommendations.** Big data analytics allow you to personalize the contents or website experiences in real time, for example, depending on gender, nationality, etc. of each consumer or from where they ended up on your site. The best-known example is probably offering tailored recommendations [15-19].

2.3. **Customization of products.** Big data can also help you to understand how others comprehend your products so that you can optimize them, or your marketing, if needed. Analysis of unstructured social media text allows you to highlight your customers' emotions and allows them to group in different geographical locations or even in different demographic groups. On the top of it, big data makes you test thousands of different types of computer-aided designs in the blink of an eye so that you can check how minor changes can be done, e.g., affecting content costs, lead times and performance. Accordingly you can change the production process [2,15].

2.4. **Risk analysis.** Success depends not only on how you run your company, and social and economic factors are also responsible for your achievements. Predictive Analytics, pointing to big data allows you to scan newspaper reports or social media feeds so that

you can maintain the pace of the latest development of the industry and its environment [15-19].

2.5. **Cost effective.** Traditionally, factories assessment reflected that a certain type of equipment is likely to become useless after few years. As a result, they need to be replaced with the latest technology equipment even though the equipment has useful life span in it. Big data tool avoids such unpractical and costly expenses. The large-scale data that they can use and their unparalleled speed figure out the faulty grid devices and predict when they will stop working. As a result, the cost-effective replacement strategy and poor downtime, the defective devices are tracked without much delay [15].

2.6. **Real-time analysis.** Previously, if business users themselves lacked the technical proficiencies required for large analysis, they had to ask their IT colleagues for help. Generally when they used to receive the requested information, it was no longer useful or correct. With the big data tools, technical teams can work with rapid algorithms to fetch the information. In other words, they can develop systems and install interactive and dynamic visibility tools that allow business users to analyze, view and make profit from data [15-19].

2.7. **Personalised healthcare.** We are living in an over-indigenous world, but one of the previous areas of healthcare still looks using a generalized approach. When someone diagnosed with cancer, they usually go through a medical treatment, and if that does not work, then doctor tries the other, etc. However, what if a cancer patient can get medicine according to his personal genes? This will result in better results, lower costs, less frustration and less fear. With human genome mapping and big data tools, this will soon be normal because everyone has one's own medical record. Hence, by finding genetic determinates, personalized medicines can be given to the patients to treat the disease and its causes [10].

2.8. **Data security.** You can map the company data with the help of big data tools, which allow you to search and analyze the threats that you might face internally. As a result, you will be able to detect the unsafe information which is potentially sensitive and also, checks whether it is stored according to the regulatory requirements or not [15].

Big data technology is coming out as a key IT component for most of the organizations and business environments and there is a growing focus on finding the business benefits of big data analytics applications to help to justify the investments in them. Big data is being used in various services and has vast range of applications in different domains as

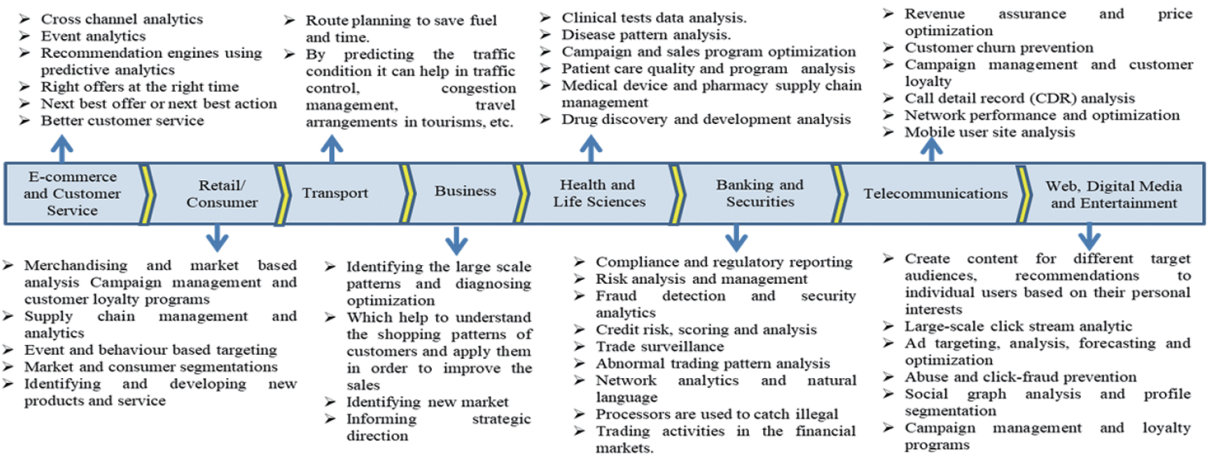

FIGURE 1. Big data applications

shown in Figure 1. Big data leads to big time benefits for the researchers and organizations. Web-based applications often encounter large amounts of data such as community computing (including social networking analysis, online communications, referral systems, reputation systems and forecasting markets), Internet text and documentation, Internet search index and other recent hotspots. There are countless sensors around us that produce small amounts of sensor data that need to be used, such as Intelligent Transportation Systems (ITS), based on analysis of a large number of complex sensor data. Large-scale e-commerce is particularly data-intensive because it involves a large number of customers and transactions and hence, has demand about it in the IT word [7,8].

3. **Big Data Analytics.** The evolution of big data leads in several organizations and enterprises in various fields. The aforementioned criteria brought the keeps to find tools and techniques for efficiently & effectively processing and analyzing the data. This type of mechanisms is called big data analytics. Big data analytics such as Hadoop is an open source framework to analyze and process big data inspired by Google's MapReduce and Google File System (GFS) papers. It is used for distribution, processing and running application for a large amount of datasets [1]. Hadoop Distributed File System (HDFS) is a core component of Hadoop and used to store input and output data. It is the block-structured files system. Currently, default block size is 128 MB which was previously 64 MB and default replication factor is 3. Block size and replication factors are configurable parameters. Hadoop MapReduce is a central module which is used to collect the data according to a query. In the MapReduce paradigm, each job has a user-defined map & reduce phase (which is processed in a completely parallel manner across a Hadoop cluster, by splitting the input data set into independent chunks and then making the data available for user consumption or additional processing) [20,21]. Initially, it was designed for a single server but later it has scaled up to thousands of machines, each offering local computation and storage. The scientific community and researchers give main focus to the advantages given by big data analytics tools while less concern to its security & privacy protection. Privacy & security protection is a set of concern which should have consider before making the big data environment. Therefore, we have highlighted some recent contributions of researchers and organizations that have enhanced and improved security & privacy of big data.

4. **Literature Review.** Probably the most challenging and related problems in big data are security and privacy protection. A large number of confidentiality areas are a group of challenges such as interacting with individuals, re-identifying attacks, potential and proven results, and economic impact. To ensure security and privacy protection, organizations should use various methods of de-identification in the big data, the most important requirements for security and privacy. Thus, security technology and other methods are always essential. Following are some potential methods and techniques used to ensure confidentiality.

Li et al. [26] proposed a new cloud architecture, MyCloud, instead of cryptographic solutions to support user-configure privacy protection in cloud environment. First, MyCloud de-privileges the cloud provider and then it enables user configured privacy protection. It also reduced the TCB size to minimize the attack surface of the cloud platform. Xu et al. [28] presented CL-PRE, it is a certificate less proxy re-encryption scheme for secure data sharing with public cloud. CL-PRE uniquely integrates identity-based public key into proxy re-encryption, eliminates the key escrow problem in traditional identity-based encryption, and does not require the use of certificates to guarantee the authenticity of public keys. Park and Lee [27] proposed a secure Hadoop architecture by adding encryption and decryption functions in HDFS, because encryption of HDFS blocks was not supported by Hadoop. Secure HDFS was implemented by adding the AES encrypt/decrypt

class to Compression Codec in Hadoop. HDFS-RSA and HDFS-Pairing [29] are the different types of integrations that are used as extensions of HDFS. These integrations can attain data confidentiality for Hadoop by providing alternatives. Novel method used [30] to encrypt file while being uploaded. Data read from a file is sent to HDFS across a buffer. In this method, an encryption is applied to the buffer's data, which is transparent to the user, before sending to an out stream to write to HDFS. In this manner, user can be unconcerned freely about the confidentiality of data. Tian [31] proposed overview of big data and discussed its security issues and summarized certain ways which improve the security of big data like security hardening methodology with attributes relation graph, attribute selection methodology, content based access control model, a scalable multidimensional anonymization approach. Also, it proposed an intelligent security model for enhancing big data security which is capable of real-time data collection and thread analysis. The model detects the thread before security intrusion in the system. The existing methods such as Password-Based Encryption (PBE) are vulnerable to brute-force attack. It is because, in the decryption process, if the key that is guessed randomly leads to some invalid-looking plain text message, it is obviously incorrect key. In the same way, the correct key can be known after numbers of hit and trials. To minimize this vulnerability, honey encryption comes in play. In [32], honey encryption mechanism is designed and implemented to the three types of private data that includes Chinese identification numbers, mobile phone numbers, and debit card passwords. The mechanism of its working is analyzed and methods to enhance the performance are proposed. Some lessons learned from honey encryption from its designing to its evaluation are also discussed.

On the basis of the above discussion, it can be inferred that in the research of the big data, the explored techniques are not sufficient as huge volume of big data is now gradually involving everywhere in various fields. Thus, it needs more privacy and security approaches and it must be explored further to identify importance of security in big data locations. Because large data security usually involves the use of big data, which implement solutions to increase the security, reliability and security of a distributed system, it needs a balance between data usage and privacy protection and data protection. The research on newly identified security approach of big data may provide security in big data environments.

5. Importance of Security in Big Data: Future Challenges.

- Greater availability and the increasing speed of Internet will contribute significantly to big data. More users are getting connected to social media such as Facebook, Instagram, and Twitter, because of faster and better Internet connectivity. The second reason for using social media networking site is that they (users) believe in actuality so we have to keep our full attention and focus on maintaining the security and privacy protection of the social networks.
- Internet of Things (IoT) and big data will play an extremely crucial role in paving the next level of generation in India because the government shifts towards digitalization like Digital India, e-governance and smart cities. Digital India initiative aims to transform the country in terms of government policies, economy, etc. into a digitally empowered society and the National Optical Fiber Network aims to connect villages with the cities through the optical fibre and provides high speed Internet & connectivity. Due to digitalization data exponentially increase, use of Hadoop and spark will grow significantly as traditional technique cannot be used efficiently and candidly to big data security perspective. This possibly provides a major impetus to big data analytics industry.
- In India, each individual must have a unique digital identify (AADHAR) card of 1.2 billion people which generated data of huge amount. Now, the government is linking this unique digital identity number to its various services/facilities like

finance, healthcare, banking, and telecommunication which results in increases of the data growth in India that creates a huge pool of data that would require big data analytics to convert the raw data into meaningful information and facilitate. That means, demand for data scientists and stronger security and privacy will continue its ascent in protecting the user's personal information.

- Further, different studies suggested that there is an exponential growth in the volume of data up to 44 ZB in the world by 2020 [2] wherein approx. 2.9 ZB increases only in India by 2020. This abrupt growth in data will boom the market and provide the abundant opportunities for midsize and small companies in big data field and also, explore this space. This will result in more companies will make use of big data for getting updates related to market trends and costumers' buying habits. Consequently, these companies will also need security to secure and protect their privacy.
- Increase of works in big data will require the use of cloud computing in hybrid manner for data collection and storing data. Then cloud computing will gain attention and conventional IT services and software development will overshadow quickly. Because of this it requires stronger security than any other securities on private servers.

6. **Discussion.** Big data is still an evolving field where much of the research has not performed yet. As of now it is handled by Hadoop, but exponential growth and spreading of data are putting even Hadoop in sort of crisis. To get most out of it in future, revolutionary technologies are needed to be devised and considerable research should be performed. From above we have concluded that security is an important aspect in the field of big data because big data includes sensitive/personal information. To achieve the better results like exciting research, marketing & better business decisions, organizations need to implement it effectively and efficiently. In this modern world of computers and Internet, technologies are exponentially evolving but everything has its pros and cons. Some evil minded people out there exploit their targets online and constantly develop new methods of doing so. These methods are often called cyber-attacks. However, security specialists are working to concoct new schemes to keep these attacks down and there is a need of finding more secure and fast methods to keep data secure. And, there is a need to focus on application security rather than device security which provides reactive and proactive protection. Experts are expectant about a new method called honey encryption [32]. It will deter hackers by serving fake data for every incorrect guess of the key code. This is a unique approach to slowing down the attackers as well as potentially burying the correct key. Apart from the above, quantum encryption methodology used in real-time information would enhance the security in a better way. This approach uses quantum principles instead of traditional methods wherein the data is transferred in a form of light (photons) called laser through a dedicated channel called optical fibre. In this approach, data is converted into special bits called QUBIT rather than normal bits (either zero or one at a time) during computation. QUBIT is a binary bit which is zero and one simultaneously at a given instance of time [33]. The private key entangled with the photons over optic fibre can solve the issues of big data security in an effective and efficient way. This private key is shared with the sending and the receiver end using a quantum channel or authenticated classical channel and if an eavesdropper tries to seal the private key in the middle of transmission it will lead to debacle. This quantum approach might have ability to work successfully now and many years ahead in the future as well.

7. **Conclusion.** Big data has created surge of opportunities for leading industries as well as government applications in recent years. Presently the data available can be used as a sample for research to tackle the big data issues expected to occur in the coming years. To understand the big data concept, and to find out the solution for the issues in the area,

it is important to know the big data characteristics as well as chronology of evolution of the same. From its inception in 1926, it has evolved exponentially in its size, volume and velocity. It is expected that the study of the concept of big data will provide better insight of the concept for researchers.

REFERENCES

[1] *https://www.youtube.com/watch?v=S89o3INzIJc_.*

[2] A. Oguntimilehin and E. O. Ademola, A review of big data management, benefits and challenges, *Journal of Emerging Trends in Computing and Information Sciences*, vol.5, pp.433-437, 2014.

[3] S. Kaisler, F. Armour, J. A. Espinosa and W. Money, Big data: Issues and challenges moving forward, *The 46th Hawaii International Conference on System Sciences*, pp.995-1003, 2013.

[4] M. Troester, *Big Data Meets Big Data Analytics*, https://eric.univ-lyon2.fr/∼ ricco/cours/slides/sources/big-data-meets-big-data-analytics-105777.pdf, 2013.

[5] Oracle, *Information Management and Big Data: A Reference Architecture*, http://www.oracle.com/technetwork/topics/entarch/articles/info-mgmt-big-data-ref-arch-1902853.pdf, 2013.

[6] C. L. P. Chen and C.-Y. Zhang, Data-intensive applications, challenges, techniques and technologies: A survey on big data, *Information Sciences*, vol.275, no.11, pp.314-347, 2014.

[7] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh and A. H. Byers, *Big Data: The Next Frontier for Innovation, Competition, and Productivity*, McKinsey Global Institute, 2012.

[8] *https://en.wikipedia.org/wiki/Big_data.*

[9] R. L. Villars, C. W. Olofson and M. Eastwood, *Big Data: What Is It and Why You Should Care?* www.idc.com, 2011.

[10] J. Roski, G. W. Bo-Linn and T. A. Andrews, Creating value in health care through big data: Opportunities and policy implications, *Health Affairs*, vol.33, no.7, pp.1115-1122, 2014.

[11] K. Shvachko, H. Kuang, S. Radia and R. Chansler, Yahoo! Sunnyvale, *The Hadoop Distributed File System*, CA, USA, 2010.

[12] B. Bornfeld and S. Rafaeli, Gamifying with badges: A big data natural experiment on stack exchange, *Frist Monday*, vol.22, no.6, 2017.

[13] A. Halavais, Home made big data challenges and opportunities for participatory social research, *Frist Monday*, vol.18, no.10, 2013.

[14] *http://www.nytimes.com/2012/03/29/technology/new-us-research-will-aim-at-flood-of-digital-data.html.*

[15] *https://dzone.com/articles/big-data-opportunties.*

[16] S. Mann, Through the glass, lightly, *IEEE Technology and Society Magazine*, vol.31, no.3, pp.10-14, 2012.

[17] F. Sestini, Collective awareness platforms: Engines for sustainability and ethics, *IEEE Technology and Society Magazine*, vol.31, no.4, pp.54-62, 2012.

[18] M. G. Michael and K. Michael, Towards a state of uberveillance, *IEEE Technology and Society Magazine*, vol.29, no.2, pp.9-16, 2010.

[19] *https://www.edureka.co/blog/10-reasons-why-big-data-analytics-is-tbesthe–career-move.*

[20] S. S. Owais and N. S. Hussein, Extract five categories CPIVW from the 9V's characteristics of the big data, *International Journal of Advanced Computer Science and Applications*, vol.7, pp.254-258, 2016.

[21] *http://data-magnum.com/how-many-vs-in-big-data-the-characteristics-that-define-big-data.*

[22] D. Agrawal, P. Bernstein, E. Bertino et al., *Challenges and Opportunities with Big Data*, https://cra.org/ccc/wp-content/uploads/sites/2/2015/05/bigdatawhitepaper.pdf, 2012.

[23] Martin Hilbert.net, *Growth of and Digitization of Global Information Storage Capacity*, www.martinhilber.net/worldinfocapacity.html, 2013.

[24] *https://www.dezyre.com/article/big-data-timeline-series-of-big-data-evolution/160.*

[25] *Apache Hadoop*, http://hadoop.apache.org/.

[26] M. Li, W. Zang, K. Bai, M. Yu and P. Liu, MyCloud: Supporting user-configured privacy protection in cloud computing, *Proc. of ACM Computer Security Applications Conference*, pp.59-68, 2013.

[27] S. Park and Y. Lee, Secure hadoop with encrypted HDFS, in *GPC 2013, LNCS 7861*, J. J. Park et al. (eds.), Springer, Berlin, Heidelberg, 2013.

[28] L. Xu, X. Wu and X. Zhang, CL-PRE: A certificateless proxy reencryption scheme for secure data sharing with public cloud, *Proc. of the ACM Symposium on Information, Computer and Communications Security*, pp.87-88, 2012.

[29] S. Ghemawat, H. Gobioff and S. Leung, The google file system, *ACM Symposium on Operating Systems Principles*, 2003.

[30] O. O'Malley, K. Zhang, S. Radia, R. Marti and C. Harrell, Hadoop security design, *Technical Report*, 2009.
[31] Y. Tian, Towards the development of best data security for big data, *Communication and Network*, Scientific Research Publishing Inc., vol.9, pp.291-301, 2017.
[32] W. Yin, J. Indulska and H. Zhou, Protecting private data by honey encryption, *Security and Communication Networks*, vol.2017, 2017.
[33] *http://gva.noekeon.org/QCandSKD/QCandSKD-introduction.html*.