Challenges, Issues, Security And Privacy Of Big Data

Taran Singh Bharati

Abstract: Data about jeans, cancers, drugs, HIV, social networking sites etc. is very huge in size and the same can be treated as big data and human is trying to decode the biological data to uncover mysteries about the biological systems. Big data is attracts people from other disciplines also. The uses and applications of Big Data are increasing day by day and it is becoming famous in data scientists biological streams. Big data keeps enormous volume and the same is being produced at fast pace from different sources. On social media thousands of posts are generated per second. Its nature, sharing, storage management, and security and privacy are some crucial issues which are taken up for consideration in this paper. The issues are thoroughly discussed and analysed.

Index Terms: Big Data, Big Data Analytics, Cloud Computing, Security, Privacy, Attacks

1 INTRODUCTION

Big data is a huge amount of data which generates heterogeneous data at high speed in usage of social media, weather forecasting, health data, industry data, space, air traffic control, science, nuclear, satellite data, etc. Data can be any kind like textual, graphical, streaming having five properties i.e. volume, veracity, velocity, and value. It is recognized by volume, velocity, variety, veracity, and value. Size is represented in Petabytes or Exabytes. Eighty percent of data which is generated by social media is unstructured data which cannot be handled by traditional software i.e. DBMS, OLAP etc. Data analysis has been named by many names in 1970 it was named by decision support system, in 1990 it became business intelligence and from 2008 it has now become data analytics. In order to process big data we require other sophisticated tools like Hadoop, R, Hive, NoSQL, search and knowledge discovery, in memory fabric, distributed file system, HDFS, Pig , data virtualization, Polybase, data integration, Sqoop, Presto, etc. The contradictions of big data are identity, transparency, and power [23]. Lambda and kappa are the big data architectures. In order to preserve the privacy, security, threats, vulnerabilities, and attacks some prevention and counter measures are needed. A system is treated secure if is access controlled, integral, authentic, and confidential. Threats and risks are exploited by adversary [24-29]. Some policies and mechanisms are framed to prevent, detect, and correct the security attacks on important data. Anti spam, antivirus, firewalls, internet security may be used to thwart the attacks. Many times data is important that its little tampering results huge wondering.

- *Dr. Taran Singh Bharati, Department of Computer Science, Jamia Millia Islamia (Central University), New Delhi, India, Emails: taran4100@gmail.com, taran_2100@yahoo.co.in*

2 GENERAL CHALLENGES AND ISSUES OF BIG DATA

There are three basic issues at storage, transportation, management, and processing level [4]:

i) Storage and Transportation Level Issue: Big data is already too big and exploding daily therefore it is very challenging to store a big data. Because there are so many big data whose sizes are in Exabytes i.e. large Hadron collider/particle physics (CRN), internet communications (CISCO), human digital universe, British library web site crawl. Since current storage technology allows around four terabytes per disk. So large number of disks would be required to store the bid data of Exabytes. It would again cause many difficulties in storage, management, and in processing. Its solution may be treated as dealing big data as much as required for operation or transportation. Similarly it stores only the results not data.

ii) Management Issue: Data is so large by nature so huge data is processed at different locations in distributive manner. So data moves from one location to another and it passes through different networks of different protocols of dealing it. We have no control on the robustness, source, platform etc.

iii) Processing Issue: since data is so big hence it would require high configuration machines. Some time data is so large that it cannot be handled by single machine because data is beyond the capacity of machines. Processing takes lot of time, sometimes it takes years to produce the result.

3 DYNAMIC DESIGN CHANGING CHALLENGES

Following characteristics must be adhering of:

i) The Size of Input / Output: Day by day the data size it increasing which cannot be handled by the traditional software.

ii) Quality and Quantity: System must ensure the availability of quantity and quality. There must be a balance between the requirements and precision.

iii) Data growth: Big data expands throughout the life time of the industry or any organization.

iv) Speed and Scale: How dada dissemination is done and on what pace.

v) Structure and unstructured: Data is both structured well formatted and well shaped and any row data in any format.
vi) Data Ownership: Some claim the owners of the data and they have the responsibility to monitor, update, and maintain the accuracy of the data for the public. So that public can trust the data.
vii) Compliance and Security: Some national and international standards, protocols and policies have to abide by the organizations.
viii) Data Granularity: Some data has some aspect and has more importance in one sense and another data has different importance in different situations.
ix) Distributed Data and Distributed Processing: Because of nature of big data, the same is contributed from several locations. The same data is processed at different locations by different processing units.
x) Compliance of Security: Some national and international standards, protocols and policies have to abide by the organizations.

4 TECHNICAL CHALLENGES OF BIG DATA ANALYSIS

Following are some challenges as scaling-extensible characteristics must be prevailed; fault tolerance-system should absorb the failures; heterogeneity-different types of documents are dealt; finding exact clue-finding the data which is more focused for business decision process; data is very vast and complex finding of relevant data is just like to jump into the ocean; combined hybrid methods-available methodologies may be combined for analyze big data sets; modelling-modelling of complex task is little challenging.

4 SECURITY AND PRIVACY OF BIG DATA

Security is protection of information assets via technology, training, and process from replay, unauthorized access, disclosure of information, inspection, and recording while privacy is the regulation or control on the personal confidential information. Things associated with security and privacy exist [13], [14], [45], [46], [47], [48]. The protection techniques are suggested as file encryption, access control, key management, logging, masking, etc. [30] [31]. In a distributed cloud environment from where one can avail on demand facilities i.e. network access at large pace hardware, software, and services from the cloud providers [2]. In distributed cloud environment some challenge [1] [2] of big data are listed below table 1.

Table1: *Security and Privacy challenges at different levels*

Network level	*Node Communication, Distributed Data and Distributed Nodes*
Authentication level	*Access Rights, Authentication Protocols, Logging, Cryptosystem*
Data level	*Distributed Data and Data Protection*
Generic types	*Traditional Tools and Use*

Privacy challenges are classified in four partitions [9], [47] as shown in figure 1. On cloud, this can be acheived by attribute based storage path, encryption, access control, etc. [20]. Integrity verification ensures that she is availing services in accordance with the singed contract.

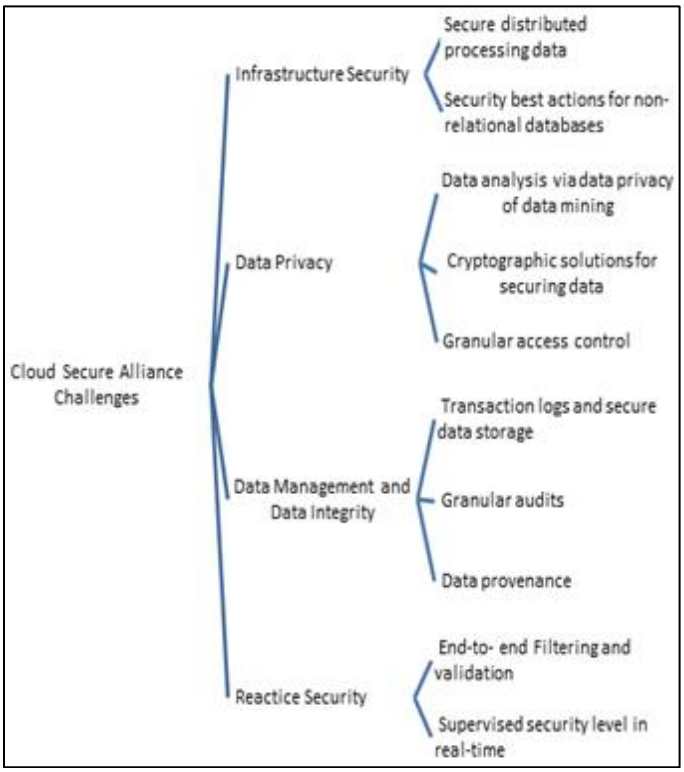


Figure1: *CSA challenges of privacy.*

More issues and challenges of security and privacy [3], [4], [5], [6], [9], [10], [12], [19] are such as semantic techniques, effective online analysis, data preparation, data sources, end-point validation/filtering, secure data storage, transmissions logs etc.

There are design issues of data input-output processes, structured versus unstructured data etc. Big Data analytics has a scaling challenge etc. Below security is proposed:

i) Hadoop Security: For authentication digital signature, SHA-256, AES, DES etc. are used. The RC6 and Kerberos protocols are used. Some securities are also listed to secure Hadoop as apache security, apache Knox, project Rhino [22].
ii) Monitoring and Auditing: A process to collect the data and check the audit records for intrusion attacks into the system.
iii) Key Management. Keys are used for transmissions. Quantum key distribution method is now supposes to be maximum secure method at low complexity for the same task [16].
iv) Cloud Security: Precious data is available at cloud and from there data is shared among the users so it requires proper authentication via encryption, decryption, compression, and login and password system.

Privacy preserving schemes named De-identification, L-diversity, K-anonymity, and T-closeness are employed. Identifier attributes, quasi identifier attributes, sensitive attributes, insensitive attributes, equivalence classes, are obvious parameters of privacy fields and they are as listed in below table2. Multiple receiver updates and conditional sharing are available to develop the gaming and collision models for thwarting attacks [19].

Table 2: *Privacy study categories*

Privacy research	
Data Clustering	*Theoretical Framework*
k-anonymity	*Differential Privacy*
l-diversity	*Differential Identifiability*
t-closeness	*Membership Privacy*

Attributes such as private attributes- sensitive keeps confidential information supposed not to be disclosed; quasi-attributes- provides indirect linkage to individuals; non-sensitive attributes- which do not reveal identity. When k-anonymity and l-diversity, and t-closeness are not successful, then we prefer (p^+, α)-sensitive k- anonymity methods [42]. We divide attributes to prevent their disclosure into different sections. In the generalization some information may be lost that is treated as a cost function [43]. The k- concealment is the secure form of k-anonymity hence k-anonymity model is k-concealment and reverse is not applicable.

6 CHALLENGES TO DATA PRIVACY

Personal information used in large data context will reveal more personal information which could lead to privacy breech of person. This information adds some value to the business which should not be used to deduce other personal information. There are legal apprehensions and legal consequences for the same by the regulatory authorities. In healthcare, patients assign rights that which information they voluntarily agree to share to agencies and which information they suppose confidential and want not to share with anyone. Some regulatory authorities are formed to protect the rights of patients' privacy [7]. There are issues for regulations like HIPAA (American Health Insurance Portability and accountability Act) and there are so many standards worldwide. Privacy fair information practices (FIP) are adopted globally with rules and regulations to handle the personal information which requires the reason for information asking and the same information should not be used for other purposes.

6.1 De-identification of Attributes

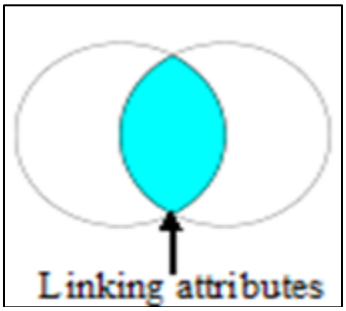
There should be a proper treatment of de-identification of person on the basis of biological parameters. The identity of individuals can be effectively known bio medically and relatively [8] [15] [17]:

i) Big Data Privacy in Data Generation: The risks are due the third party. Therefore these risks should be minimized by access restrictions and falsifying data.

ii) Big Data Privacy in Data Storage: Managing of huge data is difficult. Security protection methods are needed.

6.2 De-identification of Attributes

Information record should not involve the attribute which becomes the link to identify the other personal information of the persons [34]. For example in Indian perspective patient's data should not include PAN number, Aadhaar number, or voter identity card number because they can be used as linkage to identify the other private attributes of the person like name, age, address etc.



7 BIG DATA SECURITY TECHNIQUES

De-identifying the personal details of a person might be oral pledge, access controlled access to un-authenticated persons, encryption-decryption. Different types of encryption types like semantically secure, deterministic, order preserving, authenticated mode are suggested in literature [21]. A specialized technique called integrated rule based orientation data (iRODS) exists to ensure the privacy. Some of the techniques are as under [11]:

i) Rules and Legality: Policy level rules and regulations are framed to ensure the privacy
ii) Encryption: It is used to change the data in different forms in different places i.e. storage, communication, and computation
iii) Authentication: For controlling the access to services
iv) Metadata and tagged data: Partitioning of the information on the basis of importance and needs.
v) Unstructured Distribution: To make it difficult for malicious users.
vi) Anonymization: To use disturbances and swapping to protect the privacy personal information is de-identified and converted to secure channels.
vii) Tracking Activity: Supervision of activities and logs to control the maliciousness.

5 SECURITY INFRASTRUCTURE OF BIG DATA

In Hadoop working has a life cycle starting from data collection to delivery. It goes through following phases:

i) Scientific Data Lifecycle Management (SDLM): The security is provided to whole life cycle to keep the processing and analytics, generic data collection, filtering / classification [18].
ii) Security and Trust in Cloud Based Infrastructure: More detailed analysis like factors affecting the trustworthiness and security are considered for whole life cycle. Users know that their data is secured during data storing and processing.

iii) General Requirements for Security Infrastructure: for securing data processing the SDI and BDI must be supported in future.

6 SECURITY INFRASTRUCTURE COMPONENTS

Some of the infrastructure components in survey are proposed:

i) User or Campus Side Services: Identity management and user portals with visualization services.
ii) Federated and Delivery Infrastructure: It connects the FADI (federated access delivery infrastructure) and policy layer.
iii) Scientific and Instruments: That keeps high performance clusters of big data.
iv) Infrastructure Visualization Layer: That provides middleware and cloud/grid infrastructure services.
v) Data centres and computing resources.
vi) Network Infrastructure Layer: That represents dedicated and general internet infrastructure.

7 CONCLUSIONS

People are generating and using the big data at enormous speed in day to day life. Data flow is so fast and little regulated. So there are issues of privacy, security, identification, burglary, spreading of fake news, spreading of hates or instigation. This big data is purchased and analyzed by the business tycoons to judge trends and moods of the public. This data plays a crucial role to frame the business policies. Big data is more needy and popular to make people more aware of and alert to society. People from the biotechnology and bioinformatics analyze the big data to predict the biological disorders i.e. Cancer, HIV, Hepatitis, allergies etc. Data is gathered from heterogeneous sources and the same is utilized for analyzed for prediction. Many issues of big data exist like storage, management, security and privacy. This paper focuses and analyzes the issues in details along with special attention to security and privacy.

8 REFERENCES

[1] Inukollu VN, Arsi S, Ravuri SR. Security issues associated with big data in cloud computing. International Journal of Network Security & Its Applications. 2014 May 1; 6 3):45.

[2] Terzi DS, Terzi R, Sagiroglu S. A survey on security and privacy issues in big data. In2015 10th International Conference for Internet Technology and Secured Transactions (ICITST) 2015 Dec 14 (pp. 202-207). IEEE.

[3] Mora AC, Chen Y, Fuchs A, Lane A, Lu R, Manadhata P. Top ten big data security and privacy challenges. Cloud Security Alliance. 2012; 140.

[4] Kaisler S, Armour F, Espinosa JA, Money W. Big data: Issues and challenges moving forward. In2013 46th Hawaii International Conference on System Sciences 2013 Jan 7 (pp. 995-1004). IEEE.

[5] Matturdi B, Zhou X, Li S, Lin F. Big Data security and privacy: A review. China Communications. 2014 Apr; 11 (14):135-45.

[6] Patil HK, Seshadri R. Big data security and privacy issues in healthcare. In2014 IEEE international congress on big data 2014 Jun 27 (pp. 762-765). IEEE.

[7] Saurabh pandey, rashmi pandey, Medical (Healthcare) Big Data Security and Privacy Issues, International Journal of Scientific & Engineering Research Volume 9, Issue 2, feb-2018.

[8] Cavoukian A, Jonas J. Privacy by Design in the Age of Big Data. In Guide to Big Data Applications 2012 (pp. 29-48). Springer, Cham. Web site: www.ipc.on.ca, Privacy by Design: www.privacybydesign.ca

[9] Moura J, Serrão C. Security and privacy issues of big data. In Handbook of research on trends and future directions in big data and web intelligence 2015 (pp. 20-52). IGI Global.

[10] Lu R. Zhu H., Liu X. et al., towards efficient and privacy-preserving computing in Big Data era, IEEE Network, july-august, 2014 pp. 46-50.

[11] Gahi Y, Guennoun M, Mouftah HT. Big data analytics: Security and privacy challenges. In2016 IEEE Symposium on Computers and Communication (ISCC) 2016 Jun 27 (pp. 952-957). IEEE.

[12] Alguliyev R, Imamverdiyev Y. Big data: big promises for information security. In2014 IEEE 8th International Conference on Application of Information and Communication Technologies (AICT) 2014 Oct 15 (pp. 1-4). IEEE.

[13] Jain P, Gyanchandani M, Khare N. Big data privacy: a technological perspective and review. Journal of Big Data. 2016 Dec; 3(1):25.

[14] Perera C, Ranjan R, Wang L, Khan SU, Zomaya AY. Big data privacy in the internet of things era. IT Professional. 2015 May; 17 (3):32-9.

[15] Zhang D. Big data security and privacy protection. In8th International Conference on Management and Computer Science (ICMCS 2018) 2018 Oct 20. Atlantis Press.

[16] Thayananthan V, Albeshri A. Big data security issues based on quantum cryptography and privacy with authentication for mobile data center. Procedia Computer Science. 2015 Jan 1; 50: 149-56.

[17] Yu S. Big privacy: Challenges and opportunities of privacy study in the age of big data. IEEE access. 2016; 4: 2751-63.

[18] Demchenko Y, Ngo C, de Laat C, Membrey P, Gordijenko D. Big security for big data: Addressing security challenges for the big data infrastructure. In Workshop on Secure Data Management 2013 Aug 30 (pp. 76-94). Springer, Cham.

[19] Hu J, Vasilakos AV. Energy big data analytics and security: challenges and opportunities. IEEE Transactions on Smart Grid. 2016 Sep;7(5):2423-36.

[20] Mehmood A, Natgunanathan I, Xiang Y, Hua G, Guo S. Protection of big data privacy. IEEE access. 2016; 4:1821-34.

[21] Gadepally V, Hancock B, Kaiser B, Kepner J, Michaleas P, Varia M, Yerukhimovich A. Computing on masked data to improve the security

of big data. In2015 IEEE International Symposium on Technologies for Homeland Security (HST) 2015 Apr 14 (pp. 1-6). IEEE.
[22] Sharma PP, Navdeti CP. Securing big data hadoop: a review of security issues, threats and solution. Int. J. Comput. Sci. Inf. Technol. 2014; 5 (2):2126-31.
[23] Richards NM, King JH. Three paradoxes of big data. Stan. L. Rev. Online. 2013; 66: 41.
[24] Bharati, T. S. (2015). Enhanced Intrusion Detection System for Mobile Adhoc Networks using Mobile Agents with no Manager. International Journal of Computer Applications, 111(10).
[25] Bharati, T. S., & Kumar, R. (2015, March). Secure intrusion detection system for mobile adhoc networks. In Computing for Sustainable Global Development (INDIACom), 2015 2nd International Conference on (pp. 1257-1261). IEEE.
[26] Bharati, T. S., & Kumar, R. (2015). Intrusion Detection System for MANET using Machine Learning and State Transition Analysis. International Journal of Computer Engineering & Technology (IJCET), 6(12), 1-8.
[27] Bharati, T. S., & Kumar, R. (2016). Enhanced Key Distribution for Mobile Adhoc Networks. International Journal of Engineering Science, 6(4), 4184-4187.
[28] Bharati T. S. (2017). Agents to Secure MANETS. International Journal of Advanced Engineering and Research Development, 4(11), 1267-1273.
[29] Bharati T.S. (2018). MANETs and Its' Security. International Journal of Computer Networks and Wireless Communication, 8(4), 166-171.
[30] Jaseena KU, David JM. Issues, challenges, and solutions: big data mining. CS & IT-CSCP. 2014 Dec 27; 4 (13):131-40.
[31] Abouelmehdi K, Beni-Hessane A, Khaloufi H. Big healthcare data: preserving security and privacy. Journal of Big Data. 2018 Dec 1; 5(1):1.
[32] Li N, Li T, Venkatasubramanian S. t-closeness: Privacy beyond k-anonymity and l-diversity. In2007 IEEE 23rd International Conference on Data Engineering 2007 Apr 15 (pp. 106-115). IEEE.
[33] Sweeney L. Achieving k-anonymity privacy protection using generalization and suppression. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems. 2002 Oct;10(05):571-88.
[34] Sweeney L. k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems. 2002 Oct;10(05):557-70.
[35] Samarati P. Protecting respondents identities in microdata release. IEEE transactions on Knowledge and Data Engineering. 2001 Nov;13(6):1010-27.
[36] Truta TM, Vinay B. Privacy protection: p-sensitive k-anonymity property. In22nd International Conference on Data Engineering Workshops (ICDEW'06) 2006 (pp. 94-94). IEEE.
[37] Xiao X, Tao Y. Anatomy: Simple and effective privacy preservation. InProceedings of the 32nd international conference on Very large data bases 2006 Sep 1 (pp. 139-150). VLDB Endowment.
[38] Xiao X, Tao Y. Personalized privacy preservation. InProceedings of the 2006 ACM SIGMOD international conference on Management of data 2006 Jun 27 (pp. 229-240). ACM.
[39] Machanavajjhala A, Gehrke J, Kifer D, Venkitasubramaniam M. l-diversity: Privacy beyond k-anonymity. In22nd International Conference on Data Engineering (ICDE'06) 2006 Apr 3 (pp. 24-24). IEEE.
[40] Iyengar VS. Transforming data to satisfy privacy constraints. InProceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining 2002 Jul 23 (pp. 279-288). ACM.
[41] Bayardo RJ, Agrawal R. Data privacy through optimal k-anonymization. In21st International conference on data engineering (ICDE'05) 2005 Apr 5 (pp. 217-228). IEEE.
[42] Sun X, Wang H, Li J, Truta TM, Li P. (p+, α)-sensitive k-anonymity: A new enhanced privacy protection model. In2008 8th IEEE International Conference on Computer and Information Technology 2008 Jul 8 (pp. 59-64). IEEE.
[43] Tassa T, Mazza A, Gionis A. k-Concealment: An Alternative Model of k-Type Anonymity. Trans. Data Privacy. 2012 Apr 1;5(1):189-222.
[44] Puthal D, Sahoo BP, Mishra S, Swain S. Cloud computing features, issues, and challenges: a big picture. In2015 International Conference on Computational Intelligence and Networks 2015 Jan 12 (pp. 116-123). IEEE.
[45] Bharati T.S. (Jun, 2019). Trust Based Security of MANETs. International Journal of Innovative Technology and Exploring Engineering (IJITEE), vol. issue 8, ISSN 2278 3075, pp. 792-795.
[46] Bharati T.S. (Jun, 2019). Security and Privacy of Internet of Things. International Journal of Innovative Technology and Exploring Engineering (IJITEE) , 8(8), ISSN 2278 3075, pp 2740-2743.
[47] Bharati T.S. (Aug, 2019). Security Enhancement and Privacy Preserving of Big Data. International Journal of Innovative Technology and Exploring Engineering (IJITEE), vol.8 issue 10, ISSN 2278 3075, pp 2740-2743.
[48] Bharati T.S. (Oct, 2019). Internet of Things (IoT): Criticam Review. International Journal of Scientific & Technology Research (IJSTR), vol.8 issue 10, ISSN 2277 8616, pp 227-232.

Author's Profile

Author is B.Tech, Master of Engineering, and Ph.D. in Computer Stream from, Kanpur, Gwalior, and New Delhi respectively. He has more than 18+ years of experience. Currently he is working as senior Assistant Professor in the department of Computer Science, Jamia Millia Islamia, New Delhi.He has served at different positions in various universities and Engineering colleges. His area of interests includes Security, Theoretical Computer Science, Data Science etc.