Big Data Challenges and Issues: A Review

Akanksha Mathur(✉) and C. P. Gupta

Computer Science Department, Rajasthan Technical University, Kota, India
mathurakanksha24@gmail.com, cpgupta@rtu.ac.in

Abstract. Data is expanding immensely as well as colossally, multiplying each year. There is no denying the fact that data is and will keep on moulding our lives. Big Data can be thought of as the "development of perpetual information". Big data is pulling in technologists, researchers, and analysts in the last couple of years in different areas of large databases. Big data gathers data from multiple distributed sources in large volumes which makes it a vital issue to process data accurately for better utilization and information quality. Big data poses great challenges in many areas. The paper relates to recent findings in big data science and technology.

Keywords: Big data · Data analytics · Security and privacy · Cloud computing

1 Related Study

Big data is characterized by considering 3V's i.e. volume, variety and velocity, growing to 7v's wherein representation of data cannot be confined to conventional systems. In the 1970s, the term "Big Data" was coined, but rose in 2008. Big data defines a dataset where the data size is beyond the traditional database's capability to record, store, manage and analyze information [1]. Big data has no universally accepted definition of how large it should be for classification as big data. The data volumes are in the range of petabytes (10^{15}), exabytes (10^{18}) and beyond [2]. Data created, collected and arranged in exabyte every year. However, its creation and aggregation is quick and will approach to zeta-byte (10^{21}) in the coming years. A review of big data challenges and issues for data-centric applications expressed in type and significance of information retrieved [3]. Also, this extended too many areas further including cloud computing, IOT (internet of things), social networks, healthcare applications etc. was proven useful. Big data has advancements toward information management and handling challenges like big data analysis, knowledge diversity, knowledge extraction and reducing, integration and cleansing and several other tools for analysis and mining [4]. As business spaces are developing and there is a need to recompile the monetary framework, rethinking connections among producers, merchants, and customers of merchandise and enterprises [6].

A. P. Pandian et al. (Eds.): ICCBI 2018, LNDECT 31, pp. 446–452, 2020.
https://doi.org/10.1007/978-3-030-24643-3_53

1.1 Big Data Challenges and Issues

The most crucial part of big data is information. The appearance of any hazard indicates a need for security and protection during data transition or data storage. Classical security solutions are insufficient with reference to big data to make sure security and privacy. Hence, many security and privacy problems with big data are confidentiality, integrity, visualization and information privacy are expatiate in consequence of literature.

1.1.1 Confidentiality

Confidentiality could be a key measure to handle sensitive data, particularly, having the ability to store and process data whereas assuring confidentiality to assemble data. Confidentiality is imposing a restriction on the information against illegitimate revelation.

1.1.2 Integrity

Data integrity gives assurance against changing information by an unauthorized user. Packet sniffing, password attacks, phishing and pharming, data diddling, the man in the middle attack and session hijacking attacks are the most well-known attacks where integrity is comprised. Integrity is also maintained using data provenance, data trustiness, data loss, and data deduplication.

1.1.3 Visualization

Visualization provides a graphical representation of data that becomes easy to understand and interpret outcomes. This technique is helpful for decision makers to access, evaluate, analyze, comprehend and act on real-time data.

1.1.4 Security and Privacy

Big data contains huge amounts of individual interpersonal data that is voluminous in size and security of private data thus is the greatest challenge (Table 1).

Table 1. Big data challenges and issues

S. no.	Issues	Approaches	Description
1.	Confidentiality	Partial homomorphic encryption [8]	Proposed a secure technique for scrutiny directions, examine totally different routes avoiding data misuse of GPS information, by utilizing halfway homomorphic encryption
		Homomorphic encryption [9]	Utilized homomorphic encryption to give a convention to similitude positioning
		Authentication, Authorization, Access management (AAA) [12]	Information is encoded all through the progress and kept as plaintext
		Property-based protocol [13]	Proposed productive verification scheme for consistent information. Property-based convention for portable clients with information design system, client development procedure, and confirmation method is investigated for affirmation of classification
		Ciphertext-policy attribute-based cryptography scheme (CPABE) [14]	Generated access control scheme running in four stages: framework design, key formation, encryption, and decipherment

(*continued*)

Table 1. (*continued*)

S. no.	Issues	Approaches	Description
2.	Integrity	Feature selection along with integrity [15]	Applied a machine learning technique on a dataset for feature selection
		FPGA [16]	Used an FPGA based hardware for tamper-proof storage while maintaining confidentiality and integrity
		Tamper-resistant hardware [17]	Provide a secure token to avoid any data revelation throughout the execution of a query ensuring a closed execution environment
		Automatic analytic and on-demand data assortment methods [18]	This framework consists of on-demand automatic data collection methods
		Various approaches [19]	Various approaches like data integrity protection, digital signature, data query etc.
		Maintaining security and integrity [20]	An Approach for real-time security verification of streaming data
3.	Visualization	Various approaches to improve visualization [21]	Points out visual noise, large image perception, data loss and rate of image modification
		Toolbox for privacy preservation [22]	Enforced a toolbox for visualizing associate degreed assignment tasks supported an individual's location
4.	Security and privacy	Privacy model and techniques [23]	Explored chance of re-identification of data from various sources and model them for privacy management
		Differentially private data structures [24]	Introduced β-likeness that is a lot of informative and accessible data
		Privacy on their telecommunications platform [25]	Enforced differential privacy by maintaining the utility of the data
		Differential privacy [26]	Implement a personal querying language PINQ [27], to enhance privacy by decreasing adverse entries producing noise to produce optimum results
		PriGen framework [28]	Defines framework "PriGen" for privacy preservation in the cloud
		Reviewed privacy and protection approaches [18]	Approaches like k-anonymity, l-diversity, and t-closeness etc.
		Differentially personal learning [29]	Gathers features associated with different entities and learning useful data
		KNN formula [30]	Enforced a secure KNN within the cloud
		SNN problem [31]	Discusses attacks in existing methods for SNN, and a replacement SNN methodology that approaches to the attacks

1.2 Big Data Analytics Methods

(See Table 2).

Table 2. Big data analytical methods

S. no.	Approaches	Methodology
1.	Text Analytics [32–36]	Prediction of stock market data basically coming from financial news
		Examines text in two forms of data interpretation i.e.
		Entity Recognition (ER) and Relation Extraction (RE)
		Natural language processing technique (NLP)
		Sentiment analysis technique

(*continued*)

Table 2. (*continued*)

S. no.	Approaches	Methodology
2.	Video Analytics [37, 38]	Shared 1 s of high dimension video
		Automatic security and surveillance system
3.	Audio Analytics [39–41]	Audio analytics for treatment of specific medical conditions
		Infant cries and health status was given here
		A comprehensive review of approaches for video categorization
4.	Network Analytics [42, 43]	Large-scale graph analysis, framework supported map-reduce paradigm appeared
		Graph system on synchronous parallel model (BSP)

2 Open Research Challenges in Big Data

2.1 Security and Privacy

Many techniques developed to check personal and private information using security protocols. Still, infrastructure-based aspect for data privacy and management is an issue needed to be resolved.

2.2 Data Fusion and Visualization

Assessment of certain individual group behavior and pattern identification research carried out. An efficient storage and collection of information are required along with solving spatio-temporal problems.

2.3 Cloud Computing

On-demand services not only improve the availability of resources in the cloud environment but also works for cost reduction. Efficient management of data in storing, processing with resources utilization aspects can be considered as one aspect.

2.4 Other Social Media Related

Challenges for example online social network, data integration and performing specified operations for rumor and fake news spread can be the current state of work in big data.

3 Conclusion

Big data is a new data analysis platform that handles multidimensional information on a large-scale for data discovery and higher cognitive processes. Big Data technologies are widely used for data exploitation with the help of large - scale computing infrastructure in social big knowledge analytics tools in many areas, ranging from business intelligence to scientific exploration. This paper is a review of big data challenges, issues and methods in literature focussing on processes and tools to form the next

generation computing with big data. The paper also focuses on current research challenges in the field of big data.

References

1. Manyika, J., et al.: Big data: the next frontier for innovation, competition, and productivity, p. 1_137. McKinsey Global Institute, San Francisco (2011)
2. Kaiser, S., Armour, F., Espinosa, J.A., Money, W.: Big data: issues and challenges moving forward. In: 2013 46th Hawaii International Conference on Systems Science, pp. 995–1004 (2013)
3. Agrawal, D., Bernstein, P., Bertino, E.: Challenges and Opportunities with Big Data 2011-1 (2011)
4. Chen, M., Mao, S., Liu, Y.: Big Data: A Survey. Springer, New York (2014)
5. Chen, J., Chen, Y., Du, X., Li, C., Lu, J., Zhao, S., Zhou, X.: Big Data challenge: a data management perspective. Front. Comput. Sci. **7**(2), 157–164 (2013)
6. NIST Big Data Public Workinig Group: NIST Special Publication 1500- 4 NIST Big Data Interoperability Framework, Security and Privacy, vol. 4. September 2015
7. Murthy, P., Bharadwaj, A., Subrahmanyam, P., Roy, A., Rajan, S.: Big Data Taxonomy. Cloud Security. Alliance, no. September, p. 33 (2014)
8. Liu, A., et al.: Efficient secure similarity computation on encrypted trajectory data. In: 2015 IEEE 31st International Conference on Data Engineering (ICDE), pp. 66–77 (2015)
9. Chu, Y.W., et al.: Privacy-preserving SimRank over distributed Information network. In: 2012 IEEE 12th International Conference on Data Mining (ICDM) , pp. 840–845 (2012)
10. Bender, G., et al.: Explainable security for relational databases. In: Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data, SIGMOD 2014, pp. 1411–1422. ACM, New York (2014)
11. Meacham, A., Shasha, D.: JustMyFriends: full SQL, full transactional amenities, and access privacy. In: Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data, SIGMOD 2012, pp. 633–636. ACM, New York (2012)
12. Yang, K., Han, Q., Li, H., Zheng, K., Su, Z., Shen, X.: An efficient and fine-grained big data access control scheme with privacy-preserving policy. IEEE Internet Things J. 1–8 (2016)
13. Sudarshan, S., Jetley, R., Ramaswamy, S.: Security and privacy of big data. Studies in Big Data, pp. 121–136 (2015)
14. Jeong, Y., Shin, S.: An efficient authentication scheme to protect user privacy in seamless big data services. Wirel. Pers. Commun. **86**(1), 7–19 (2015)
15. Xiao, H., et al.: Is feature selection secure against training data poisoning? In: Proceedings of the 32nd International Conference on Machine Learning (ICML 2015), pp. 1689–1698 (2015)
16. Arasu, A., et al.: Secure database-as-a-service with cipherbase. In: Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data, SIGMOD 2013, pp. 1033–1036. ACM, New York (2013)
17. Lallali, S., et al.: A secure search engine for the personal cloud. In: Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, SIGMOD 2015, pp. 1445–1450. ACM, New York (2015)
18. Xu, L., Shi, W.: Security theories and practices for big data. Big Data Concepts, Theories, and Applications, pp. 157–192 (2016)

19. Gao, Y., Fu, X., Luo, B., Du, X., Guizani, M.: Handle a framework for investigating data leakage attacks in hadoop. In: 2015 IEEE Global Communications Conference (GLOBECOM), San Diego, CA, pp. 1–6 (2015)
20. Puthal, D., Nepal, S., Ranjan, R., Chen, J.: A dynamic key length based approach for real-time security verification of big sensing data Stream. Lecture Notes in Computer Science, pp. 93–108 (2015)
21. Gorodov, E.Y., Gubarev, V.V.: Analytical review of data visualization methods in application to big data. J. Electr. Comput. Eng. 22 (2013)
22. To, H., et al.: PrivGeoCrowd: a toolbox for studying private spatial Crowdsourcing. In: 2015 IEEE 31st International Conference on Data Engineering (ICDE), pp. 1404–1407 (2015)
23. Lu, R., et al.: Toward efficient and privacy-preserving computing in big data era. IEEE Netw. **28**(4), 46–50 (2014)
24. Cao, J., Karras, P.: Publishing microdata with a robust privacy guarantee. Proc. VLDB Endow. **5**(11), 1388–1399 (2012)
25. Hu, X., et al.: Differential privacy in telco big data platform. Proc. VLDB Endow. **8**(12), 1692–1703 (2015)
26. Proserpio, D., et al.: Calibrating data to sensitivity in private data analysis: a platform for differentially private analysis of weighted dataset
27. McSherry, F.D.: Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In: Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data, pp. 19–30. ACM (2009)
28. Rahman, F., Ahamed, S., Yang, J., Wang, Q.: PriGen: a generic framework to preserve privacy of healthcare data in the cloud. In: Inclusive Society: Health and Wellbeing in the Community, and Care at Home, pp. 77–85 (2013)
29. Jain, P., Thakurta, A.: Differentially private learning with kernels. In: Proceedings of the 30th International Conference on Machine Learning (ICML 2013), pp. 118–126 (2013)
30. Elmehdwi, Y., et al.: Secure k-nearest neighbor query over encrypted data in outsourced environments. In: 2014 IEEE 30th International Conference on Data Engineering (ICDE), pp. 664–675 (2014)
31. Yao, B., et al.: Secure the nearest neighbor revisited. In: 2013 IEEE 29th International Conference on Data Engineering (ICDE), pp. 733–744 (2013)
32. Chung, W.: BizPro: intelligence factors from textual news articles. Int. J. Inf. Manag. **34**(2), 272–284 (2014)
33. Jiang, J.: Information extraction from text. In: Aggarwal, C.C., Zhai, C. (eds.) Mining text Data, pp. 11–41. Springer, New York (2012)
34. Hahn, U., Mani, I.: The challenges of automatic summarization. Computer **33**(11), 29–36 (2000)
35. Liu, B.: Sentiment analysis and opinion mining. Synth. Lect. Hum. Lang. Technol. **5**(1), 1–167 (2012)
36. Feldman, R.: Techniques and applications for sentiment analysis. Commun. ACM **56**(4), 82–89 (2013)
37. Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., et al.: Big data: the frontier for innovation, competition, and productivity. McKinsey Global Institute (2011). http://www.citeulike.org/group/18242/article/9341321
38. Shockley, R., Smart, J., Romero-Morales, D., Tufano, P.: Gains with "big data," according to a conducted on behalf of SAP (2012)
39. Patil, H.A.: "Crybaby": assess neonatal health status from an infant's cry. In: Neustein, A. (ed.) Advances in Speech Recognition, pp. 323–348. Springer, New York (2010)

40. Hirschberg, J., Hjalmarsson, A., Elhadad, N.: “You’re as sick as you sound”: using computational approaches to gauge illness and recovery. In: Neustein, A. (ed.) Advances in Speech Recognition, pp. 305–322. Springer, New York (2010)
41. Hu, W., Xie, N., Li, L., Zeng, X., Maybank, S.: A survey on visual content-based video indexing and retrieval. IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. **41**(6), 797–819 (2011)
42. Elser, B., Montresor, A.: An evaluation study of big data frameworks for graph processing. In: Proceedings of IEEE International Conference on Big Data, pp. 60–67. IEEE (2013)
43. Valiant, L.G.: A bridging model for parallelcomputation. Commun. ACM **33**(8), 103–111 (1990). https://doi.org/10.1145/79173.79181
44. Bertino, E., Ferrari, E.: Big Data Security and Privacy. Springer, Berlin (2018)