Issues and Challenges in Big Data: A Survey

Ripon Patgiri(✉)

Department of Computer Science and Engineering,
National Institute of Technology Silchar, Assam 788010, India
ripon@cse.nits.ac.in
http://cse.nits.ac.in/rp/

Abstract. Undoubtedly, the Big Data is the most promising technology to serve an organization in a better way. It provides an organized way to think about data, whatever the data size is, and whatever the data type is. Moreover, the Big Data provides a platform to make decisions, and to analyze future possibilities using the past and present data. The Big Data technology eases the large dataset to store, process and manage. The Big Data is the most fashionable trendsetter in the world of computing i.e., the most popular buzzword around the globe upon which the future of the most of IT industries depends on it. This paper presents a study report on numerous research issues and challenges of Big Data which is employed in very large dataset. This paper uncovers the nuts and bolts of Big Data. This study report provides rich insight on the Big Data.

Keywords: Big Data · Big Data survey · Big Data paradigm
Issues and challenges · Big Data analytics · Big Data security
Healthcare

1 Introduction

The Big Data is not only the most promising technology, but also a social necessity to reclaim their lifestyle. Facebook, for instance. Therefore, data are increasing at an exponential pace. The Big Data is a high volume of data, and Seife quotes this data-ism as, "Data-ism is very much a conventional business book, full of anecdotes, mini-profiles and aphorisms that grow ever less compelling [24]". The Big Data concerns about assigning worth to those high volume of data. The Big Data comprises of unstructured, semi-structured and structured data. The Big Data is proven as a game changer in many data-intensive field. The Big Data enhance the decision making process automatically [16]. *A wrong decision can destroy an organization.* That's why, Big Data Analytics evolves to assist and guide the decision-making process. The data-driven decision-making process is the most crucial and critical part of an organization to make a move [26]. However, the Big Data face the "curse of dimensionality" issue. Many new services evolve based on Big Data, namely, Big Data as a Service, Big Data Security as a Service, Big Health Data as a Service, and Big Data Analytics as a Service.

A. Negi et al. (Eds.): ICDCIT 2018, LNCS 10722, pp. 295–300, 2018.
https://doi.org/10.1007/978-3-319-72344-0_25

Besides, the Big Data has nothing untouched area, namely, engineering, science, government, economy, and environment.

Interestingly, Einav et al. [11] describes the role of Big Data in economic growth. McAfee et al. [21] emphasize more on the generation of business revenue using Big Data. Therefore, there are numerous field which smoothens by Big Data technology, namely, eHealth, Wearable technology, customization of Medicine, Internet of Things (IoT), customer analytics solution, surveillance system, transport system, Digital experience solution and Power/Energy. Bourne et al. [5] quote as, "the research community must find more efficient models for storing, organizing and accessing biomedical data". The biocuration requires access to Big Data by both biomedical and biological discoveries [14]. Therefore, big interdisciplinary data are also growing. These technologies produce an enormous volume of data. The well-known data-intensive fields are, namely, GIS, weather, earthquake, gnome, ocean, soil, oil, and drone. Therefore, computer scientists, physicists, economists, mathematicians, bio-informaticists, and sociologists use Big Data for various purposes. The government and private sectors are the key areas to use Big Data analytics. Thus, a massive amount of data is spawned excessively with the course of time, and therefore, these data are cumbersome to store, process, visualize and analyze the data in conventional ways. Most of these data are from different fields, and these are unstructured. Therefore, interdisciplinary research on Big Data is prominent challenge and opportunity nowadays.

2 Issues and Challenges in Big Data Analytics

Big Data analytics is a method of logical analysis on a very large dataset. There are numerous purposes of practicing Big Data analytics (BDA) and results enormous possibilities in research. For example, BDA leads to better healthcare [22] which is the most prominent research challenge nowadays. BDA is used in decision support system in healthcare for a better outcome, prevention, low-cost solution and early detection of an event. The BDA dig into the data for new facts or insight. The first key challenge of BDA is the monitoring real-time events with high volume, and to explore the available high volume of data. The second key challenge of BDA is the prediction of future based on Big Data using machine learning algorithms. The BDA is deployed in the business organization to know the behavior of their business and to know the future using their own data set. The third key challenge is the decision making process using a very large data set which requires BDA. The key challenges of the BDA is to learn the decisions and make a right decision based on huge volume of the dataset. The BDA suggests a good solution among many solutions in large dataset. The fourth key challenge is the diagnostic analytics of BDA to discover about past performance/events. Fifth, uncover the hidden patterns of the past events is also a challenge. The BDA recognizes the behavior of users and data to detect anomalies. Finally, the most promising challenge is the anomaly detection, intrusion detection, and fraud detection over a very large dataset. This is a tough challenge to achieve.

3 Issues in Big Data Security

Protecting and securing data is the top priority of the Big Data paradigm. Moreover, Big Data is used to detect security threats. On the contrary, the Big Data is also used to detect the breach of a system to attack/test. The Big Data eases the capability of securing data, protecting data and ensuring the privacy of data. The Big Data analytics address data security, technology to keep customers data private, provenance, data transparency, performance benchmarking, data and system interoperability. There are two aspects of Big Data security (BDS), namely, Big Data for security and Security for Big Data. Moreover, secure infrastructure, secure data management, data privacy, and real-time security are some example of BDS [12]. The requirement of BDS is confidentiality, authenticity, integrity and availability (-service should not down due to DDOS) [3]. Moreover, BDS reduces the breach of risk sensitive data. Cloud Security Alliance (CSA) categorizes BDS as infrastructure security, data protection, data management and reactive security [2,13]. However, CSA enlisted top ten challenges in Big Data Security [13], namely, (a) secure computations in distributed programming frameworks, (b) security best practices for non-relational data stores, (c) secure data storage and transaction logs, (d) endpoint input validation/filtering, (e) real-time security/compliance monitoring, (f) Scalable and composable privacy-preserving data mining and analytics, (g) cryptographically enforced access control and secure communication, (h) granular access control, (i) granular audits, and (j) data provenance.

The BDA is used to detect of anomaly, intrusion, fraud, and advanced persistent threats (APT) [6]. It is impossible to spell-check the large size security analysis in a conventional way. The security of Big Data data has been achieved by deploying BDA on the large sized logs, system events, network traffic, website traffic, security information and event management (SIEM) alert, cyber attack patterns, business processes and other information sources. Besides, access control of the billion users is the perplex job [3]. It is an open challenge to protect data from malicious attackers. The diversity of data sources, data formats, streaming of data and infrastructures can lead to security vulnerabilities.

4 Open Challenges

The challenges of Big Data are outlined below-

(a) The Big Data is really big enough to transmit data from one source to another. (b) A large dataset is difficult to visualize, very tough to mine a meaningful information, and perplex to manage these data. (c) These huge sets of data consist of structured data, unstructured data, and semi-structured data. The key issue is the various sources of data, which forms various kinds of data. Storing these data heterogeneity itself a great challenge. (d) The real-time Big Data processing is a big challenge. (e) It is a challenge to make a correct decision using the large dataset. (f) The efficient visualization of data is an open challenge [15]. (g) The "pay-as-you-go" model helps in decreasing the costs of

users. The Big Data as a Service and Big Data Analytics as a Service significantly reduces the costs. However, it is still a challenge in the Big Data paradigm for lowering the cost. (h) The most prominent issues in load balancing are heterogeneity, scalability, consistency, and adaptability. (i) The fault-tolerance system is the most cumbersome for administrator and fault cannot be obviated easily. The fault-tolerance model is implemented by RAID, replication, erasure coding, de-duplication, and journaling. (j) The Big Data technology requires auto-scalability with dynamic data size. The scalability is the big issue in the Big Data. The designing infinite scalability is the biggest challenge. (k) Another research challenge is the achieving high performance using the low-cost commodity hardware. (l) The key issue is dynamic volume and the technology requires to adjust itself to cope up with the ever changing environment. (m) The prominent issue is to design an automatic failover mechanism to ensure high availability. (n) The bandwidth is not unlimited to transfer data in a real scenario, thus, reducing bandwidth consumption a challenge. (o) Another issue is the ameliorating the throughput significantly. (p) The performance of a file system depends on the how minimal network traffic has generated. A fine tuning of network traffic is required to excel in performance in data-intensive computing. (q) The disaster recovery and management are the big issue and the big challenge for all time. The Disaster Recovery is the most cardinal part of data storage system. Disaster Recovery as a Service or Recovery as a Service is the most prominent emerging cloud model in disaster recovery. (r) The data acquisition is an issue of Big Data. Data does not come automatically, but user makes bigger database size. Either data is collected explicitly or implicitly, database size grows continuously. (s) The data curation of Big Data concerns with data reuse, data discovery, and data preservation, such that the value of the data is maintained over time [1,7]. Especially, the data curation in Big Data becomes more complex due to high volume.

5 Issues and Challenges in Big Data Applications

The Big Data is very complex to deploy in real system due to mammoth sized data, and moreover, it is continuously monitored, processed, and visualized. However, the data-intensive fields use Big Data technology to enhance their revenue and performance, like Biomedical engineering. Undoubtedly, the Big Data is a good choice for Biomedical engineering due to the massive amount of data to be analyzed [8]. The Big Biomedical Data Engineering (BBDE) requires huge storage spaces, processing capacities, visualization and analysis. The article [4] ask a question- "why do we write?". The assumed environment may differ, but the answer is similar. However, the biomedical engineering requires the data to write, so that someone will use in future to study the diseases in the curing process. The answer converges with article [4]. Big Data Analytics (BDA) is a merger of Big Data and Analytics [9]. The analytics means the logical method of analysis. BDA provides a platform to discover the hidden jewels from data.

6 Discussion, and Future Direction

Big Data technology is developed to serve the billions of clients for the purposes of the generating revenue. The future of Big Data targets more on Interdisciplinary computing [23]. Lynch [18] quote as, "If data cannot survive in the short term, it is pointless to talk about long-term use". Bourne et al. [5] call for a more efficient way of storing, managing and processing the Medical data. Landhuis [17] reports the Neuroscience is another emerging field for Big Data because neuron size of any species is very big to store and process. Marx et al. [19] report that the Big Data is required in the cancer study. Nature [10] editorial quote as, "Health professionals will confront more data than do those in finance". Topol [25] quote as, "a massive, open, online medicine resource would help to quickly identify the genetic cause of the disorder". The Big Data is used to enhance the healthcare process [27]. Moreover, the NASA process petabytes of Climate data [20]. Another future agenda is the real-time processing of the monster size data [8]. The real-time Big Data processing is a very complex process. A strong programming paradigm is required to process real-time Big Data efficiently in the scale of infinite (Exabyte or beyond). Moreover, the Big Data span from pernicious project to constructive project. All people on the earth will engage with Big Data from 2020 and onward either directly or indirectly.

7 Conclusions

We have exposed the issues and challenges of Big Data. The key issues of Big Data are volume, velocity and variety. Moreover, security, privacy, adaptability, fault-tolerance, consistency, data curation, data acquisition, network traffic, bandwidth and latency, performance, scalability, load balancing are also some prominent issues of Big Data technology. The BDS and BDA also play vital role which is the prominent issues for many organizations for many years. We have also discussed that the direction of Big Data is moving towards "Interdisciplinary Big Data Computing" and "Very Big Data". The scope of Big Data is not limited to engineering, Science, environment, economic, biology, and agriculture. For example, the Big Data can be used in the medical domain, like cancer treatment, and brain analysis.

References

1. Abe, A.: Curating and mining (big) data. In: 2013 IEEE 13th International Conference on Data Mining Workshops, pp. 664–671 (2013)
2. Alguliyev, R., Imamverdiyev, Y.: Big data: big promises for information security. In: IEEE 8th International Conference on Application of Information and Communication Technologies (AICT 2014), pp. 1–4 (2014)
3. Bertino, E.: Big data - security and privacy. In: 2015 IEEE International Congress on Big Data, pp. 757–761 (2015)

4. Bonenfant, M., Desai, B.C., Desai, D., Fung, B.C.M., Ozsu, M.T., Ullman, J.D.: Panel: the state of data: invited paper from panelists. In: Proceedings of the 20th International Database Engineering & Applications Symposium, pp. 2–11 (2016)
5. Bourne, P.E., Lorsch, J.R., Green, E.D.: Perspective: sustaining the big-data ecosystem. Nature **527**(7576), S16–S17 (2015)
6. Cardenas, A.A., Manadhata, P.K., Rajan, S.P.: Big data analytics for security. IEEE Secur. Priv. **11**(6), 74–76 (2013)
7. Chen, C.P., Zhang, C.-Y.: Data-intensive applications, challenges, techniques and technologies: a survey on big data. Inf. Sci. **275**(2014), 314–347 (2014)
8. Cuzzocrea, A., Sacca, D., Ullman, J.D.: Big data: a research agenda. In: Proceedings of the 17th International Database Engineering & Applications Symposium, pp. 198–203 (2013)
9. Desai, B.C.: Technological singularities. In: Proceedings of the 19th International Database Engineering & Applications Symposium, pp. 10–22 (2015)
10. Editorial: The power of big data must be harnessed for medical progress. Nature, **539**(7630), 467468 (2016)
11. Einav, L., Levin, J.: Economics in the age of big data. Science **346**(6210), 1243089 (2014)
12. Fang, W., Wen, X.Z., Zheng, Y., Zhou, M.: A survey of big data security and privacy preserving. IETE Technical Review, pp. 1–17 (2016)
13. Big Data Working Group: Expanded top ten big data security and privacy challenges. Cloud Security Alliance, pp. 1–39, April 2013
14. Howe, D., Costanzo, M., Fey, P., Gojobori, T., Hannick, L., Hide, W., Hill, D.P., Kania, R., Schaeffer, M., St Pierre, S., et al.: Big data: the future of biocuration. Nature **455**(7209), 47–50 (2008)
15. Jagadish, H.V., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J.M., Ramakrishnan, R., Shahabi, C.: Big data and its technical challenges. Commun. ACM **57**(7), 86–94 (2014)
16. Labrinidis, A., Jagadish, H.V.: Challenges and opportunities with big data. Proc. VLDB Endowment **5**(12), 2032–2033 (2012)
17. Landhuis, E.: Neuroscience: big brain, big data. Nature **541**(7638), 559–561 (2017)
18. Lynch, C.: Big data: How do your data grow? Nature **455**(7209), 28–29 (2008)
19. Marx, V.: Biology: the big challenges of big data. Nature **498**(7453), 255260 (2013)
20. Mattmann, C.A.: Computing: a vision for data science. Nature **493**(7433), 473475 (2013)
21. McAfee, A., Brynjolfsson, E., Davenport, T.H., Patil, D., Barton, D.: Big data: the management revolution. Harvard Bus. Rev. **90**(10), 61–67 (2012)
22. Schadt, E.E.: The changing privacy landscape in the era of big data. Mol. Syst. Biol. **8**(612), 1–3 (2012)
23. Schadt, E.E., Linderman, M.D., Sorenson, J., Lee, L., Nolan, G.P.: Computational solutions to large-scale data management and analysis. Nat. Rev. Genet. **11**(9), 647657 (2010)
24. Seife, C.: Big data: the revolution is digitized. Nature **518**(7540), 480–481 (2015)
25. Topol, E.J.: The big medical data miss: challenges in establishing an open medical resource. Nat. Rev. Genet. **16**(5), 253254 (2015)
26. Wang, H., Xu, Z., Fujita, H., Liu, S.: Towards felicitous decision making: an overview on challenges and trends of big data. Inf. Sci. **367–368**(2016), 747–765 (2016)
27. Wang, Y., Hajli, N.: Exploring the path to big data analytics success in healthcare. J. Bus. Res. **70**(2017), 287–299 (2017)