Challenges for Big Data Analytics: A Review

Dharmender Kumar
Guru Jambheshwar University of Science & Technology, Hisar, Haryana, India

Narender Kumar
Guru Jambheshwar University of Science & Technology, Hisar, Haryana, India

Abstract - The data is very essential part of the information society that the world has turned into. Data has become very important for information systems which empower business, industry, organization and individual. The generation of vast amount of records by various information systems day by day has given birth to an era of big data. The term "Big Data" expresses a very massive size of datasets such that traditional database software tools are incapable to store, manage and analyze them efficiently. Hence the current data analytics fail to process Big Data. The main challenges for the researchers are to develop an efficient platform for big data analysis with high performance and to design appropriate big data mining algorithms. These challenges are discussed deeply in the survey, which starts with an introduction to big data, review of the literature on big data analytics and challenges for big data analytics.

Keywords:Big Data, Big Data Analytics, Data Mining, Large Datasets.

I. INTRODUCTION

The data have been generating at a very rapid rate in the world and with each passing day a very large amount is being added to itto grow further. The emerged technologies, like the digital and social media and internet of things, are empowering big data to extend more. According to estimation more than 92% of new information has been produced by digital media devices in 2002 [1]. So it is more difficult and challenging to analyze the big data to find useful things, even when we are using very fast computer systems, because the data is being created at a very high rate. The size of big data is not the only concern for researchers but the present data analytics mechanism cannot be applied to it directly. So it becomes very challenging for existing data processing and data management tools to process such a large collection of data. This is the reason, the most updated information systems or applications are not capable to cope with big data. In addition, traditional data mining methods or data analytics may not be applied directly to big data because they were developed for centralized data analysis and it is not possible to load all the data into a single machine, in big data era.

In recent years, several studies based upon data analytics and frameworks have been presented, with their pluses and minuses, while there are two perspectives this survey provides as a review on big data analytics: a) platforms and framework perspective and b) data mining algorithms perspective.

The term "Big Data" is used to identify the datasets that cannot be managed by existing data mining software tools because of two main reasons: large size and complexity. Big Data mining is the process of extracting useful information or knowledge from these large datasets. Doug Laney talked about three dimensions of big data also called as 3 V's in Big Data management [3]:

Volume: volume means not only the amount of data but unexceptionally more data with ever increasing, making it difficult to handle for existing data analytics.

Variety: data is of different types like structured data, unstructured data, semi-structured data and more complex structured data.

Velocity: data are created, stored, analyzed and visualized at a very high speed.

Now, there are more V's: [2]

Variability: it is the user's way to evaluate the data and data dimensions resulting from multiple disparate data types and sources.

Value: big data is valuable from business point of view to understand customer better, optimizing business processes.

Veracity: it refers to reliability of the data source and its context, the analysis of the data is based on the reliability.

Validity: validity specifiesthat the data correct and accurate for the purpose. Clearly valid data should yield accurate results for business.

Venue: data would be coming from different platforms, owners and in different formatting requirements.

Vocabulary: the origin, syntax and structure of data are different due to semantics, data models schema etc.

Vagueness: jumble over the big data interpretation.

This paper is begins with an introduction to big data, the research in frameworks, the research in mining algorithms, main challenges in the era of big data and then the conclusions and future trends.

In the next section, the researches in frameworks and mining algorithms are explained as literature review. This review is done from two perspectives: 1) research in platforms and frameworks and 2) research in mining algorithms.

II. LITERATURE REVIEW

The data analysis is the process of accumulating, arranging and analyzing huge data sets in order to find meaningful information and knowledge patterns. The unique features of big data are size, complex, diverse, inadequate, incoherent, high dimensionality, faulty and noisy. Hence analysis of big data is challenging as well very important, because big data analytics can help in business decisions by identifying useful information. It also makes organizations to understand information contained within the data better. Because of so many unique features big data may contain unclear and irregular data. For instance, a user may have multiple phone numbers, which may reduce the accuracy of the mining results. Hence there are some new challenges for data analytics arise, such as privacy,fault tolerance, storage, security, and quality of data [11]

In this next section research done in frameworks and platforms in past few years are studied.

2.1 Research in Big data frameworks and platforms

The large datasets cannot be handled, at once, by existing computer systems so it becomes very important to design an efficient data analytics framework for big data.

The Apache Hadoop [9] software library is an open-source software framework for distributed storage and processing of large data sets using MapReduce programming across clusters. It allows applications to work a large set of independent computers on large size of data.

To develop a general model for computation in big data analytics, Huai et al. [12] presented a matrix model called DOT. This model comprised of three matrices for data set (D), concurrent data processing operations (O), and data transformations (T) respectively. This model enforces data independent relationship to achieve fault tolerance and scalability. In this model big data analytics job is represented by DOT blocks, and the results are collected and transformed to a computer node.

Rusu F et al. [13] presented generalized linear aggregates distributed engine (GLADE)for efficient big data analytics. The GLADE is multi-leveldata analytics system that provides a distributed runtime environment where GLAs are executed in parallel.

In [14],Herodotou et al. provided solutions to select targets according to user needs and system workloads in virtual reality (VR) environments. In their study, Starfish, a self-tuning analytics system for big data analysis was presented. The performance of this system seems good due to self-adjustment according to user needs.

The study by Demchenko et al [15] was from the perspectives of data centric architecture and operational models to present a big data architecture framework (BDAF) which includes: big data infrastructure, big data analytics, data structures and models, big data lifecycle management, and big data security.

Ye et al. [16] presented architecture called cloud-based big data mining and analyzing services platform (CBDMASP). This architecture integrates R to rich data statistical and analyticfunctions. The architecture includes four layers: infrastructurelayer, virtualization layer, dataset processing layer andservices layer.

In the research work done by Wu et al [17], a theorem,called HACE, to characterize the big data characteristics and a big data processing model, were proposed. This model included data accessing sources,mining and analysis tools, user interest modeling and security and privacy considerations. The study concluded that in big data mining the design of big data analytics platform is very challenging and difficult.

In [18] Laurila etal. addressed that in data analytics, data collected from mobile devices, privacy should be the important factor to be considered. So in big data mining security is an essential challenge to handle.

In the study by Demirkan and Delen[19] it was addressed that a service-oriented decision support system (SODSS) for big data analytics is a major trend. They proposed a conceptual framework for DSS in cloud and addressed time and cost in delivering big data analytics as service are the major challenges.

In [20] Talia indicated the requirement of new models, tools, andtechnologies to implement dynamic dataanalysis algorithms, where data analytics software, data analytics platform and infrastructure can be provided as service in cloud-based data analytics.

Lu Ret al. [21]addressed efficient and privacy preserving computing as new challenges in big data analytics. They presented an efficient and privacy-preserving cosinesimilarity computing protocol by analyzing general architecture of big data analytics which included multi-source big data collecting, distributed big data storing and processing.

In [22],Cuzzocrea et al. discussed about research issues and achievements in big data analytics. They extended the discussion,to pointout open problems in big multidimensional data analytics, to some open issues like filtering uncorrelated data and possible research leads for these problems.

Zhang and Huang [23], in their study, proposed a 5Ws model for big data analysis in order to explore new big data approach. The model represents types of data, origin of data, causes of data, time of data occurrence,receiver of data and means to transfer the data.

Chandarana P et al. [24] explored the characteristics and issues in big data analytics by comparing Hadoop [9], Storm and Drill frameworks. This study highlighted big data workloads, privacy, security, datamanagement and technology, are the key challenges for the design of big data platform.

[25] Hu et al. in their study discussed about a big data chain comprised of data generation, data acquisition, data storage, and dataanalytics. In addition to it,in the study, they presented a layered decomposition big data system having infrastructure, computing, and application layers.

2.2 Research in Big data mining algorithms

The data mining algorithms are very important in big data analytics. Some important factors which are dependent upon the big data mining algorithms are cost, storage requirements, accuracy etc. In this section a brief discussion for big data analytics from the perspective of mining algorithms is done.

Shirkhorshidi etal. [26],presented a review on the progress of clustering algorithms in big data mining and found that dealing with hugeamount of data is very challenging for them. Further to it, the multiple machine clustering can be the solution provided the complexity of the same it is properly addressed.

Xu etal [27] presented a solution, CloudVista, for clustering big data. This solution can work on entire data set using parallel processing framework.

Cui et al. [28] presented a multiple species flocking (MSF) approach to achieve a better performance in clustering algorithms. The simulation results obtained from the evaluation, of 100 news articles from internet, are quite promising.

Feldman et al. [29] used merge and reduce approach in which by steaming parallel coresets. This approach reduced the update time significantly for existing clustering algorithms.

Tekin et al [30] addressed the problem of decentralized big data classification and presented an online classification algorithm for the same. In this study heterogeneous distributed learners cooperatively learn the classifications functions in order to improve accuracy.

In [31], Rebentrost et al. implemented asupport vectormachine on a quantum computer for classification. The time complexity of the quantum algorithm is logarithmic in the size dimensions and training data.

In the studies [35] and [37] the performance of the mining algorithm was improved significantly. These approaches used map-reduce solution which will be the future of mining applications on cloud platform [34, 4]. The study of [37] also included users' interest in the mining process.

In another study,by Ku-Mahamudet al [32], big data clustering was done on grid computing platform. In this approach a modified ant-based clustering algorithm is used, where each ant is used in parallel computing environment, on random machines, in the grid.

III. CHALLENGES IN BIG DATA ANALYTICS

a) Heterogeneity and Incompleteness: -

The machine analysis algorithms are incapable to deal with heterogeneity and unable to understand the depth of natural language due to variation in expression. And there is great involvement of heterogeneity when natural language is used in human interactions. So it is very challenging to structure data appropriately before data analysis.

b) Volume:-

Undoubtedly, it is quite challenging to deal with rapidly growing volume of data. Computer resources have certain limit in the extension of their processing power to deal with this challenge. So this is a challenging issue t deal with.

c) Timeliness:-

It is quite obvious that processing time is directly related to volume of data. So another important issue is the speed. So the design of a system that can efficiently process a given data set in a reasonable time is challenging.

d) Privacy:-

Another big concern is the privacy of data in big data analytics. There are certain rules and strict laws to deal with data that is related to public. Big data can be comprised of personal data, data linked through multiple resources, health records etc. So this issue must be addressed properly in big data analytics.

IV. CONCLUSIONS

This paper highlights the concept of big data analytics from the framework and platform perspective as well as the mining perspectives. The challenges which are open for researchers, from both the perspectives, regarding quality, security and computation of data are then discussed. The possible research trends derived from this study are given below.

1. Distributed parallel computing is one of the important future trend to make data analytics work for big data. Management of computation resources including speed of data analysis, scheduling of tasks in cloud based environment is another future trend.

2. Making data mining algorithms work on parallel computing environment for big data analytics will be a future research trend.

3. Representation of big data analytics results will also be a future trend because it will decide the how data analytics will work in real world practically.

4. Analyzing big data exists over social network has become a promising research issue.

5. The security and privacy issues will also be the future research trends in big data analytics.

REFERENCES

[1] Lyman P, Varian "H. How much information 2003?" Tech. Rep, 2004. [Online]. Available: http://http://www.groups.ischool.berkeley.edu/archive/how-much-info-2003/printable_report.pdf.

[2] Zhang J, Huang ML. 5Ws model for big data analysis and visualization. In: Proceedings of the International Conferenceon Computational Science and Engineering, 2013. pp 1021–1028.

[3] Albert Bifet, "Mining Big Data in Real Time," Informatica 37, pp. 15–20, Dec 2012.

[4] Huang JW, Lin SC, Chen MS, "DPSP: Distributed progressive sequential pattern mining on the cloud", Proceedings of the Advances in Knowledge Discovery and Data Mining, vol. 6119, pp 27–34, 2010.

[5] FiratTekiner and John A. Keane, "Big Data Framework," IEEE International Conference on Systems, Man, and Cybernetics, pp. 1494-1499, Jan 2013.

[6] BenediktElser and Alberto Montresor, " An Evaluation of Big Data framework for Graph Processing," IEEE International Conference on Big Data, pp. 60-67, Mar 2013.

[7] SharanjitKaur and DhritiKhanna, "Scalable Clustering Using PACT Programming Model," IEEE 12th International Conference on Data Mining Workshops, pp. 424-430, May 2012.

[8] FiratTekiner and John A. Keane, "Big Data Framework," IEEE International Conference on Systems, Man, and Cybernetics, pp. 1494-1499, Jan 2013.

[9] ApacheHadoop, June 22, 2014. [Online]. Available: http://hadoop.apache.org.

[10] Lee J, Hong S, Lee JH. "An efficient prediction for heavy rain from big weather data using genetic algorithm,"International Conference on Ubiquitous Information Management and Communication, pp 25–27, 2014.

[11] Katal A, Wazid M, Goudar R, "Big data: issues, challenges, tools and good practices", Proceedings of the International Conference on Contemporary Computing, pp 404–409, 2013.

[12] Huai Y, Lee R, Zhang S, Xia CH, Zhang X. "DOT: a matrix model for analyzing, optimizing and deploying software for big data analytics in distributed systems", Proceedings of the ACM Symposium on Cloud Computing, pp 4-14, 2011.

[13] Rusu F, Dobra A. "GLADE: a scalable framework for efficient analytics", Proceedings of LADIS Workshop held in conjunction with VLDB, pp 1–6, 2012.

[14] Wonner J, Grosjean J, Capobianco A, Bechmann D "Starfish: a selection technique for dense virtual environments",Proceedings of the ACM Symposium on Virtual Reality Software and Technology, 2012. pp 101–104, 2012.

[15] Demchenko Y, de Laat C, Membrey P "Defining architecture components of the big data ecosystem",Proceedings of the International Conference on Collaboration Technologies and Systems,pp 104–112, 2014.

[16] Ye F, Wang ZJ, Zhou FC, Wang YP, Zhou YC "Cloud-based big data mining and analyzing services platform integrating R", Proceedings of the International Conference on Advanced Cloud and Big Data, pp. 147–151, 2013.

[17] Wu X, Zhu X, Wu G-Q, Ding W, " Data mining with big data", IEEE Transaction Knowledge Data Engineering,. pp. 97–107, 2014.

[18] Laurila JK, Gatica-Perez D, Aad I, Blom J, Bornet O, Do T, Dousse O, Eberle J, Miettinen M. "The mobile data challenge: big data for mobile computing research", Proceedings of the Mobile Data Challenge by Nokia Workshop, pp. 1–8, 2012

[19] Demirkan H, Delen D, "Leveraging the capabilities of service-oriented decision support systems: putting analytics and big data in cloud" Decision Support System, pp. 412–421, 2013.
[20] Talia D, "Clouds for scalable big data analytics", IEEE Computer Society, pp. 98–101, 2013.
[21] Lu R, Zhu H, Liu X, Liu JK, Shao J, "Toward efficient and privacy-preserving computing in big data era", IEEE Network;28(4):46–50, 2014.
[22] Cuzzocrea A, Song IY, Davis KC, "Analytics over large-scale multidimensional data: The big data revolution!", Proceedings of the ACM International Workshop on Data Warehousing and OLAP, pp. 101–104, 2011
[23] Zhang J, Huang ML, "5Ws model for big data analysis and visualization", Proceedings of the International Conference on Computational Science and Engineering, pp 1021–1028, 2013.
[24] Chandarana P, Vijayalakshmi M, "Big data analytics frameworks", Proceedings of the International Conference on Circuits, Systems, Communication and Information Technology Applications, pp. 430–434, 2014.
[25] Hu H, Wen Y, Chua T-S, Li X, "Toward scalable systems for big data analytics: a technology tutorial", IEEE Access, pp. 652–687, 2014.
[26] Shirkhorshidi AS, Aghabozorgi SR, Teh YW, Herawan T. "Big data clustering: a review", Proceedings of the International Conference on Computational Science and Its Applications, pp. 707–720, 2014.
[27] Xu H, Li Z, Guo S, Chen K, "Cloudvista: interactive and economical visual cluster analysis for big data in the cloud", Proc VLDB Endowment, pp. 1886–1889, 2012.
[28] Cui X, Gao J, Potok TE, "A flocking based algorithm for document clustering analysis", J SystArchit, pp. 505–515, 2006.
[29] Feldman D, Schmidt M, Sohler C, "Turning big data into tiny data: Constant-size coresets for k-means, pcaandprojective clustering", Proceedings of the ACM-SIAM Symposium on Discrete Algorithms, pp. 1434–1453, 2013.
[30] Tekin C, van der Schaar M, "Distributed online big data classification using context information", Proceedings of the Allerton Conference on Communication, Control, and Computing, pp. 1435–1442, 2013.
[31] Rebentrost P, Mohseni M, Lloyd S. Quantum support vector machine for big feature and big data classification.
[32] CoRR, vol. abs/1307.0471, 2014. [Online]. Available: http://dblp.uni-trier.de/db/journals/corr/corr1307html#RebentrostML13.
[33] Ku-Mahamud KR, "Big data clustering using grid computing and ant-based algorithm",Proceedings of the International Conference on Computing and Informatics, pp. 6–14, 2013.
[34] Yang L, Shi Z, Xu L, Liang F, Kirsh I, "DH-TRIE frequent pattern mining on hadoop using JPA",Proceedings of the International Conference on Granular Computing,pp 875–878, 2011.
[35] Lin MY, Lee PY, Hsueh SC, "Apriori-based frequent itemset mining algorithms on mapreduce", Proceedings of the International Conference on Ubiquitous Information Management and Communication, pp 76–78, 2012.
[36] Riondato M, DeBrabant JA, Fonseca R, Upfal E "PARMA: a parallel randomized algorithm for approximate association rules mining in mapreduce", Proceedings of the ACM International Conference on Information and Knowledge Management, pp 85–94, 2012.
[37] Leung CS, MacKinnon R, Jiang F, "Reducing the search space for big data mining for interesting patterns from uncertain data",Proceedings of the International Congress on Big Data, pp 315–322, 2014.