Research in Big Data and Analytics: An Overview

Lekha R. Nair
Research Scholar,
BITS Pilani, Dubai Campus
Dubai, UAE

Sujala D. Shetty, Ph.D.
Assistant Professor,
BITS Pilani, Dubai Campus
Dubai, UAE

ABSTRACT
Big Data Analytics has been gaining much focus of attention lately as researchers from industry and academia are trying to effectively extract and employ all possible knowledge from the overwhelming amount of data generated and received. Traditional data analytic methods stumble in dealing with the wide variety of data that comes in huge volumes in a short period of time, demanding a paradigm shift in storage, processing and analysis of Big Data. Owing to its significance, several agencies including U.S. government have released huge funds for research in Big Data and allied fields in recent years. This paper presents a brief overview of research progress in various areas associated to Big Data Processing and Analytics and conclude with a discussion on research directions in the same area.

General Terms
Data Analytics, Machine Intelligence, High Performance Computing, Data Mining, Big Data

Keywords
Big Data Analytics, Big Data Processing, Big Data Research

1. INTRODUCTION
In today's digitally connected world, every single thing can be considered as of generating data. This data that is getting added to the existing ocean of Big Data formed from myriad sources like web logs, smart phones, social network sites, satellite images, human genomics, customer transactions, astronomical and biological records, poses immense opportunities as well as challenges for researchers to tackle and provide beneficial outcome.

The term Big Data is not just about the magnitude of data that comes in the range of petabytes or zeta bytes; rather it is more about the capability to handle huge amounts of data.

The University of California, Berkeley defines "Big data is when the normal application of current technology doesn't enable users to obtain timely, cost-effective, and quality answers to data-driven questions" [1]. Big data is featured mainly by 3 V's - Volume, Velocity and Variety where in massive volume of data of orders of exabyte or zettabyte possessing huge variety as text files, images, documents, videos, log files coming in complex forms as structured, unstructured and semi structured formats, having varying velocity requirements as real time or near real time processing along with batch processing. More V's are added to Big Data dimensionality such as value that can be derived from the Big Data and veracity which defines the understandability of Big Data. Storage and manipulation of big data and the transformation of big data to knowledge are the major issues associated with Big Data. It is often thought that the huge volume of Big Data ensures that valuable knowledge is buried underneath which needs to be found, but analysts cannot just intuit for themselves for the valuable content of the data [2].Traditional data analytic techniques which works on structured data of low volume found to be inefficient to handle the variety and complexity offered by the big data which is mostly unstructured or semi-structured and that comes in huge volume. A comparison of Big Data analytics with traditional analytics is given in table 1.

Table 1. Traditional Data Analytics vs. Big Data Analytics

	Traditional Analytics	**Big Data Analytics**
Data Sources	Trusted homogenous sources providing structured and static data	Heterogeneous sources providing unstructured/ semi structured and streaming data
Data Storage	Isolated proprietary servers	Public/Private/ Hybrid Cloud
Database Technology	Relational data stores	NoSQL data stores
Data Processing	Centralized Architecture	Distributed Architecture
Analytics	On previously collected data	Need for real time analytics

Big data analytics can provide new insights to every field. It can lead to efficient detection or prediction of new scientific theories, customer behavior, social phenomena, weather patterns, economic conditions [3] etc., which in turn aids in better decision support. Scientific community is counting on Big Data analytics as the patterns unfold from this could never have been found by other means before.

2. BIG DATA RESEARCH
Research in Big Data progresses in various dimensions including effective capture of data, finding novel storage solutions and retrieval techniques for this massive data, exploring cost effective solutions for transportation of data from storage grid to processing grid, implementation of performance improved big data processing frameworks with reduced latency and increased throughput, formulation of scalable machine learning and data mining algorithms for precise learning and accurate predictions, developing new visualization techniques, adopting tight security and privacy preservation strategies and benchmarking of Big Data analytic system. Figure 1 gives an overview of Big Data analytics flow. Moreover, much attention has been given to Big Data analytics in cloud environment. Social network sourced big data analytics is also an active area of research to find cutting edge trends.

2.1 Big Data Collection
There is no shortage of Big Data samples; a single sequenced human genome is about 140 gigabytes in size [4]; in astronomy, new telescopes are generating over a petabyte of data per day [5]. While Facebook claims an average of 829 billion daily active users in June 2014, it is not hard to

imagine how huge will be the amount of data generated by this social network site in terms of photos uploaded besides likes and comments posted. With innumerable sources of Big Data available, focus should be on reaping its benefits without further delay.

2.2 Big Data Storage

With data explosion, Big Data storage systems need to have large and growing capacity, high band width, ability to handle fluctuating load characteristics, reduced I/O path delays, techniques to deal with semi structured and unstructured data without compromising on reliability and security.

Storage strategy can be a centralized one which is simpler and having less communication cost, or a distributed one which is more reliable and extendable. Though Google File System and Hadoop Distributed File System are so popular in Big Data scenario, according to [6], these have issues like Small Files Problem as they are mainly meant to handle large files and not the smaller internet files that is having large meta data and whose access frequency is higher, which leads to performance degradation. Also for smaller files, file fragmentation may lead to wastage of disk space and creating links for each small file may lead to network delays. Load balancing in distributed strategy is also an issue which requires consistency in replicated data. De-duplication which deletes redundant data for storage capacity optimization is also very popular.

Since storage drives are slow, to perform analytics on Big Data, in-memory analytics are favored as it probes the data is stored in RAM which speeds up the process [7].

Spark [8] introduced Resilient Distributed Dataset (RDD), the in-memory distributed partitions that are kept persistently in the memory of cluster nodes, to eliminate hard disk drive input/output latency. Iterative applications are benefited as it can cache intermediate data which can be used in further iterations.

Cloud storage [9] is a popular storage medium for the big data though it is not referring to a separate medium but to the fact that data is stored offsite and is accessed online. Many organizations are now migrating their big data to clouds like Amazon AWS, IBM SmartCloud, and Windows Azure for storage. Cloud provides hierarchical tiered storage mechanism that makes use of flash arrays/solid state technology, hard disks and tapes, and efficient storage management software where the medium of storage is chosen on priority basis based on varying requirements as latency, cost, energy efficiency, capacity and reliability, while these underlying things are hidden from the end user. Movement of data to the cloud and maintenance of privacy and security in the cloud are still challenging issues in consuming cloud storage

2.3 Big Database Technology

In [10], Sam Madden discusses how databases are effective in solving Big Data problems. Traditional databases, mainly the relational databases, lack scalability, elasticity, fault tolerance and flexibility and hence seem to be a bad choice in distributed systems. Since Big Data demands scale out systems for collecting and interpreting data, scale out data stores commonly referred as NoSQL (Not Only SQL) systems are given much significance compared to relational databases. NoSQL databases are offering flexible schema and elasticity and it can be implemented using cheap commodity hardware, though there is a compromise in ACID (Atomicity, Consistency, Isolation, and Durability) transactions. Based on data model, NoSQL databases may be either Key-value, Column –oriented, Document stores or graph databases. For Key value database, a value will be associated with a key and it is having high query speed and concurrency compared to relational databases. DynamoDB, MemCache DB, Redis, Voldemort are examples of Key-value stores.

OrientDB, Allegro and Virtuso are graph databases which are useful in dealing with data where relationships play an important role as in social networking. In the graph, nodes represent entities and edges represent relationships. Document databases are similar to key-value stores, but the value or data is stored in documents in some form of markup language like JSON or XML. MongoDB and CouchDB are document data stores. BigTable, HBase, HadoopDB, and Cassandra are Column oriented data stores which are tabular in nature. A comparison of NoSQL databases is given in [11].

In [12], Tim and Beth explains about new database architectures in Big Data and in [13], models and languages for big data querying was discussed and insights to designing a querying engine has been given.

2.4 Big Data Processing

MapReduce [14], introduced by Google, is the programming model that provides abstraction from underlying hardware and facilitates parallel programming and execution on multiple clusters. Hadoop [15] is the open source implementation of MapReduce, is a popular big data processing engine. It is originally a batch processing system and it prefers throughput, reliability and scalability over execution time or latency, which introduces certain bottle necks. It also lacks the support of real time processing that deals with frequently changing dynamic data [16]. Various optimizations on the open source implementation of MapReduce to tune it in accordance with the requirement of big data analytics and to cater efficient data centric computation have been proposed. MapReduce++ [17] suggests the optimization of response time by utilizing a scheduling strategy similar to the Shortest-Job-First by calculating the task's time cost and executing smaller ones first. Stubby [18] is a workflow optimizer that generates MapReduce workflows, which is extensible and transformation based, but has the drawback of not considering all types of transformation.

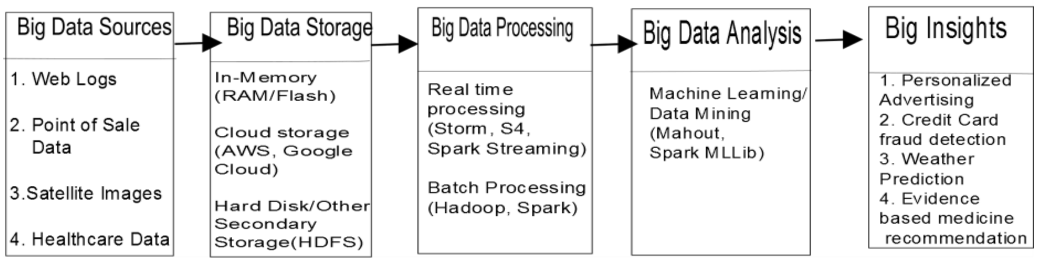

Fig 1: The small picture of Big Data analytic flow

Starfish [19], which is built on Hadoop, provides a self-tuning system for the analytics of Big Data, by automatically adapting to user requirements and system workloads so as to yield better performance. Radoop [20], an integration of Hadoop with the data miner tool Rapidminer, scales well with increasing data size and combines the advantages of both.

Sailfish [21] another MapReduce framework which uses an abstraction termed I-files for data aggregation and transporting data from map tasks to reduce tasks and is performing 20% to 5 times faster compared to Hadoop.

Twister [22] extends MapReduce by adding iterative support for MapReduce tasks which helps in continuous data processing. Haloop [23] also works in a similar manner and found to have reduced execution time compared to Hadoop. Still both suffer from limitations like impossibility of triggering on new data, high response time and batch processing model which makes them unsuitable for real time stream processing.

Twitter Storm [24] works well for real time streaming data where data being given as infinite set of tuples to the system. They are transferred to processing nodes where it is consumed and results are published or more data streams are produced for further processing/aggregation. Storm is able to process million tuples of data per node per second. Spark streaming is also popularly used in handling streaming data

2.5 Big Data Transportation

Though big data analytics can be effectively performed in the cloud environment, transfer of the massive data set to the cloud seems to be a challenge. L. Zhang [25] proposed an online cost minimizing approach to upload this data to a distant cloud. Two online algorithms for optimization of the choice of the data center when moving geo-dispersed data as well as the routes to transmit data to that particular data center is also discussed in the paper.

2.6 Big Data Analytics

Data analytics is considerably difficult compared to data collection and storage. Development of scalable and parallel machine learning algorithms for online analytics has been a serious challenge [16]. Apache Mahout Project provides several parallel machine learning algorithms tuned for MapReduce execution but mostly meant for batch processing. Hence they don’t directly support iterative or online stream processing though the available parallelization features can be tuned to online processing.

In [26], authors introduces certain principles that can be used to design a flexible data analytic pipeline and establish architectures for effective discovery of knowledge from Big Data.

Existing analytic frameworks are mostly transaction based which have been effectively used in business applications like customer segmentation and marketing, management of financial and accounting activities, while the perspective is being shifted towards ecosystem based analytic frameworks which focus on integrated analysis of the less structured environments, compared to the isolated transaction analysis [27].

In [28], authors presents a HACE theorem that models Big Data characteristics and proposes a Big Data processing model from a data mining viewpoint. As per the paper, Big Data conceptual framework consists of three tiers, with tier 1 dealing with data access and computing, concern of tier II being data privacy and domain knowledge and tier III being Big Data mining/machine learning algorithms. Platforms built for large-scale data analytics can only handle a fraction of machine learning algorithms at scale and better systems support for already established machine learning use cases remains as an open research question [29]. Traditional data mining algorithms require loading of entire data in the main memory for performing mining, but for big data it is expensive to move data across various locations. Domain knowledge of applications is also essential as data privacy and data sharing mechanisms can be different based on the nature and requirement of the applications. Mining complex semantic relationship from Big Data improves performance of applications including search engines and recommendation systems and gives insight into various social phenomena but it has become a great challenge due to the heterogeneity and huge volume of the data.

2.7 Security and Privacy Preservation

Privacy preservation of data in the cloud has been a matter of concern as the available strategies are not sufficient to prevent leakage of sensitive personal information. As organizations can benefit from the analysis of sensitive data like health records and financial transactional records, failures of traditional privacy protection measures in cloud can be utilized by malicious users to divulge data privacy, which may severely affect social reputation or may result in financial loss for the owners of data. Data anonymization techniques, where identity or other sensitive data of owners is concealed, are widely used for privacy preservation and several anonymization algorithms have been proposed. But with big data trends, these anonymization algorithms fail to anonymize such huge data set and researchers are trying to improve the scalability issues in anonymization of large data sets [30]. A two phase scalable top down specialization approach is proposed in [31] to anonymize large scale data set making use of MapReduce on cloud.

Differential privacy paradigm is in use recently that protects leakage of private data while execution of queries on data. In [32] an algorithm called DiffMR is proposed, which process top-k query over MapReduce framework while maintaining differential privacy.

US government had released two reports in May 2014 – White House report and PCAST report that is mainly related to privacy related issues of Big Data. These reports declines encryption as a perfect solution for privacy preservation and points to the inadequacies of data anonymization and de-identification techniques. It gives policy recommendation for responsible and accountable usage and disclosure of Big Data [33]

2.8 Visualization and Benchmarking

In the big data era, development of novel visualization techniques are in progress and [34] gives a detailed review of the popular visual analytic systems. In [35] main challenges associated with Big data visualizations and ways to avoid the same are discussed. Big Data Visualization methods like Tree maps, Circle packing, Sunburst, Circular Network Diagram, Parallel coordinates and Streamgraphs have been analyzed and compared in this paper. An interactive visualization tool named TopicFlow has been described in [36] which is meant to be used with twitter that aligns and displays similar topics from time slices over a period of time.

Designing benchmark suite to evaluate the performance of Big Data systems is also a matter of concern to researchers and in [37], analysis of some of the available benchmark suites like HiBench and ICTbench are given.

2.9 Big Data in Cloud Environment

Issues and challenges associated with Big Data processing in a cloud environment from the perspectives of user, data and hardware have been given extensively in [9]. Zimmermann et al. discusses on a service oriented enterprise classification model for Big Data in the cloud environment [38]. Ji et al. [39] specifies key issues in Big Data processing systems running in the cloud environment. Here various approaches regarding big data platform, distributed file system for storage, and optimization of MapReduce were discussed. An analytical model based on queuing theory was proposed in [40] to achieve elasticity for MapReduce jobs running in cloud clusters by giving an estimate of resources required. Simulation of the model was done but its experimental validation was not done.

For the deployment of dynamic MapReduce clusters in multi cluster systems, Ghit [41] in his research paper has proposed a resource management system that ensures provisioning and scheduling decisions in accordance with workload characteristics.

2.10 Social Network Sourced Big Data

Social networks like Facebook and Twitter are leading producers of big data. Social Network topology has strong impact on physical technological networks as most of the traffic is contributed by these social network sites and related ones. In [42], some methodologies and cases were discussed in which the application of social network analysis for designing technological networks has been done and vice-versa. A user can be a member of several networks simultaneously and these networks form a composite social network where the user can exhibit different behavior in different networks. At the same time the user may share some common latent interests across these networks as well. E. Zhong [43] proposed a model for adaptive transfer of knowledge from composite social networks to predict human behavior which can be utilized in social marketing, service recommendations and personalization.

Leveraging social network paradigm for deriving knowledge from big data by using personal ad hoc clouds of participants in social networks to tackle big Data processing challenges has been presented in [44].

3. SOCIAL ISSUES IN BIG DATA

Now the organizations that employ Big Data analytics have competitive advantages over those who don't. Cloud based analytics are very popular due to scalability and cost effectiveness, but many smaller companies are staying away due to lack of network bandwidth and money.

Though immense data from human physical activity like phone calls, location, health records, transactions has been collected, according to [45], the moral and ethical aspects of the use of Big Data is vague. The ways of analyzing this data and the choices that are made to act upon this data has serious human consequences, though the data is neutral. Thus the same census data can be analyzed and used by the government to decide upon effective funding policies or to decide on exploiting communal sentiments on selecting candidates for the coming elections. This points to the need of proper monitoring on the usage of Big Data as well as the type of data that companies could legitimately collect. Big Data analytics has been proven beneficial to big companies, but smaller companies who previously used to find it expensive and complex are also now shifting to utilize its merits [46].

4. CONCLUSION AND FUTURE SCOPE

Research in various facets of Big Data is accelerating to catch up with the Big Data flow rate. Although this paper gives some flavor of Big Data research progress, several research questions in this area are still open for future research works. There is a growing need to address issues like data and tool interoperability for various data formats and tools as well as integration of various Big Data analytic frameworks. Security and privacy preservation of sensitive information shared, especially in cloud, is another zone of great concern for researchers. Streaming data acquisition is demanding efficient real time analytic frameworks for the future. Real time security monitoring is a challenging task that needs much research attention. Defining the appropriate programming abstractions for domain specific applications and formulation of optimized balanced scheduling strategies for large scale data sets necessitates further improvements. Research scope exists in development of scalable and parallelizable machine learning algorithms for complex Big Data where Deep Learning is also gaining much attention. Further research should aim at finding effective and practical solutions to these problems. It is definite that the valuable treasures of knowledge hidden under the deep layers of Big Data Ocean will continue to captivate researchers for years to come.

5. REFERENCES

[1] T. Kraska, "Finding the Needle in the Big Data Systems Haystack," IEEE Internet Computing, vol. 17, no. 1, pp. 84-86, 2013.

[2] F. Shull, "Getting an Intuition for Big Data," IEEE Software, vol. 30, no. 4, pp. 3-6, 2013.

[3] C. Jayalath, J. Stephen and P. Eugster, "From the Cloud to the Atmosphere: Running MapReduce across Data Centers," IEEE Transactions on Computers, vol. 63, no. 1, pp. 74-87, 2014.

[4] V. Marx, "The Big Challenges of Big Data," Nature, vol. 498, no. 7453, pp. 255-260, 2013.

[5] H. S. Francis J. Alexander, "Big Data," Computing in Science and Engineering, vol. 13, no. 6, pp. 10-12, 2011.

[6] F. X. Zhang Xiaoxue, "Survey of Research on Big Data Storage," in IEEE International Symposium on Distributed Computing and Applications to Business, Engineering & Science, 2013.

[7] L. Garber, "Using in-memory analytics to quickly crunch big data," Computer, vol. 45, no. 10, pp. 16-18, 2012.

[8] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker and I. Stoica, "Spark: Cluster Computing with Working Sets," in USENIX conference on Hot topics in cloud computing, 2010.

[9] R. Branch et al., "Cloud Computing and Big Data: A Review of Current Service Models and Hardware Perspectives," Journal of Software Engineering and Applications, vol. 7, pp. 686-693, 2014.

[10] S. Madden, "From Databases to Big Data," IEEE Internet Computing, vol. 14, no. 6, pp. 4-6, 2012.

[11] H. Jing et al., "Survey on NoSQL database," in International Conference on Pervasive Computing and Applications, 2011.

[12] K. Tim and B. Trushkowsky, "The New Database Architectures," IEEE internet computing, vol. 17, no. 3, pp. 72-76, 2013.

[13] B. Novikov, N. Vassilieva and A. Yarygina, "Querying Big Data," in International Conference on Computer Systems and Technologies, 2012.

[14] J. Dean and S. Ghemawat, "Mapreduce: Simplified Data Processing on Large Clusters," Communications of the ACM, vol. 51, no. 1, pp. 107-113, 2008.

[15] T. White, "Hadoop: The Definitive Guide, 3rd Edition", O'Reilly Media, California, 2012.

[16] Osman, M. El-Refaey and A. Elnaggar, "Towards Real-Time Analytics in the Cloud," in IEEE Ninth World Congress on Services, 2013.

[17] Z. Guigang, L. Chao, Z. Yong and C. Xing, "MapReduce++: Efficient Processing of MapReduce Jobs in the Cloud," Journal of Computational Information Systems, vol. 8, no. 14, pp. 5757-5764, 2012.

[18] L. Harold, H. Herodotos and B. Shivnath, "Stubby: a transformation-based optimizer for MapReduce workflows," Proceedings of the VLDB Endowment, vol. 5, no. 11, pp. 1196-1207 , 2012.

[19] H. Herodotou, H. Lim, G. Luo, N. Borisov and L. Dong, "Starfish: A Self-tuning System for Big Data Analytics," in 5th Biennial Conference on Innovative Data Systems Research (CIDR '11), California, 2011.

[20] Z. Prekopcs´ak, G. Makrai, T. Henk and C. G´asp´ar-Papanek, "Radoop: Analyzing Big Data with RapidMiner and Hadoop," in 2nd RapidMiner Community Meeting and Conference (RCOMM 2011), 2011.

[21] S. Rao et al., "Sailfish: A Framework for Large Scale Data Processing," in Proceedings of the Third ACM Symposium on Cloud Computing, 2012.

[22] J. Ekanayake, "Twister: A Runtime for Iterative Mapreduce," in 19th ACM International Symposium on High Performance Distributed Computing, 2010.

[23] B. Yingyi, B. Howe, M. Balazinska and M. D. Ernst, "HaLoop: Efficient iterative data processing on large clusters," in VLDB Endowment, 2010.

[24] W. Yang, X. Liu, L. Zhang and L. T. Yang, "Big Data Real-time Processing Based on Storm," in 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications, 2013.

[25] C. W. Linquan Zhang, Z. Li, C. Guo, M. Chen and F. C. Lau, "Moving Big Data to The Cloud:An Online Cost-Minimizing Approach," IEEE Journal on Selected Areas In Communications, vol. 31, no. 12, pp. 2710-2721, 2013.

[26] B. Edmon and J. Horey, "Design principles for effective knowledge discovery from big data," in Joint Working IEEE/IFIP Conference onSoftware Architecture (WICSA) and European Conference on Software Architecture (ECSA), 2012.

[27] Z. Daniel and R. Lusch, "Big Data Analytics: Perspective Shifting from Transactions to Ecosystems," IEEE Intelligent Systems, vol. 28, no. 2, pp. 2-5, 2013.

[28] X. Wu, X. Zhu, W. Gong-Qing and W. Ding, "Data Mining with Big Data," IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 1, pp. 97-107, 2014.

[29] T. Condie, P. Mineiro, N. Polyzotis and M. Weimer, "Machine Learning for Big Data," in ACM SIGMOD International Conference on Management of Data, New York, USA, 2013.

[30] K. LeFevre, D. J. DeWitt and R. Ramakrishnan, "Workload-aware Anonymization Techniques for Large-scale Datasets," ACM Transactions on Database Systems, vol. 33, no. 3, pp. 17:1-17:47, 2008.

[31] Z. Xuyun, L. T. Yang, C. Liu and J. Chen, "A scalable two-phase top-down specialization approach for data anonymization using mapreduce on cloud," IEEE Transactions on Parallel and Distributed Systems, vol. 25, no. 2, pp. 363-373, 2014.

[32] X. Han, M. Wang, X. Zhang and X. Meng, "Differentially Private Top-k Query over Map-Reduce," in Fourth ACM international workshop on Cloud data management, 2012.

[33] B. M. Gaff, H. E. Sussman and J. Geetter, "Privacy and Big Data," IEEE Computer, vol. 47, no. 6, pp. 7-9, 2014.

[34] L. Zhang, "Visual analytics for the big data era—A comparative review of state-of-the-art commercial systems," in IEEE Conference on Visual Analytics Science and Technology, 2012.

[35] G. E. Yur'evich and V. V. Gubarev, "Analytical review of data visualization methods in application to big data," Journal of Electrical and Computer Engineering, vol. 2013, pp. 1-7, 2013.

[36] S. Malik et al., "TopicFlow: Visualizing Topic Alignment of Twitter Data Over Time," in Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2013.

[37] W. Xiong, "A Characterization of Big Data Benchmarks," in IEEE International Conference on Big Data, 2013.

[38] Zimmermann, M. Pretz, G. Zimmermann, D. G, Firesmith and I. Petrov, "Towards Service-oriented Enterprise Architectures for Big Data Applications in the Cloud," in IEEE International Enterprise Distributed Object Computing Conference Workshops, 2013.

[39] Ji, L. Yu, Q. Wenming, A. Uchechukwu and L. Keqiu, "Big Data Processing in Cloud Computing Environments," in International Symposium on Pervasive Systems, Algorithms and Networks, 2012.

[40] K. Salah and J. M. A. Calero, "Achieving Elasticity for Cloud MapReduce Jobs," in IEEE 2nd International Conference on Cloud Networking, San Francisco, 2013.

[41] B. Ghit, A. Losup and D. Epema, "Towards an Optimized Big Data Processing System," in 13th IEEE/ACM International Symposium on Cluster, Cloud, and Grid Computing, 2013.

[42] C. K. Cheng, M. Chiang and H. V. Poor, "From Technological Networks to Social Networks," IEEE Journal on Selected Areas in Communications, vol. 31, no. 9, pp. 548-572, 2013.

[43] E. Zhong, W. Fan, J. W. L. Xiao and Y. Li, "ComSoc: Adaptive Transfer of User Behaviors over Composite Social Network," in 18th ACM SIGKDD international conference on Knowledge discovery and data mining, 2012.

[44] W. Tan, M. Blake, I. Saleh and S. Dustdar, "Social-Network- Sourced Big Data Analytics," IEEE Internet Computing, vol. 17, no. 5, pp. 62-69, 2013.

[45] G. Booch, "The Human and Ethical Aspects of Big Data," IEEE Software, vol. 31, no. 1, pp. 20-22, 2014.