BIG DATA ANALYTICS - AN OVERVIEW

[1] V. Vaneeswari, [2] S.Ranichandra, [3] S. Selvakumari
[1][2][3] Assistant Professor,
Department of Computer Science
Dhanalakshmi Srinivasan College of Arts and Science for Women (Autonomous),
Perambalur.

ABSTRACT

Big Data Analytics has been in advance more attention recently since researchers in business and academic world are trying to successfully mine and use all possible knowledge from the vast amount of data generated and obtained. Demanding a paradigm shift in the storage, processing and analysis of Big Data, traditional data analysis methods stumble upon large amounts of data in a short period of time. Because of its importance, the U.S. Many agencies, including the government, have in recent years released large funds for research in Big Data and related fields. This gives a concise summary of investigate growth in various areas related to big data processing and analysis and terminate with a discussion of research guidelines in the similar areas.

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

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

1. INTRODUCTION

Nowadays digitally related world, every sole thing can be measured as of produced data. This data, which is incorporated into large data analyzes generated from numerous sources such as networks, telephones, different sites, satellites, human genetics, customer interaction, redundancy and records, poses great opportunities and challenges for researchers to deal with and deliver favorable results.

The term Big Data is not limited to the amount of data that falls within the range of beta bytes; slightly about the convenience of handling large amounts of data.

University of California, Berkeley defines ig Big data as regular use of current capability makes it impossible for users to get relevant, profitable, and better answers to data-driven questions - [1]. Large data is obtained mainly through 3 Vs - Speed, Volume and Variety, where large amounts of data in jetabytes are available in images, documents, text files, videos, log files in complex formats, unconfigured and semi-configured formats, with different speed rations for real-time or block processing Are close to real-time processing. Additional V is added to large data components such as 3V that describe the potential value of Big Data and the comprehensibility of Big Data. Large data handling and storage and changes of large data to knowledge are key issues associated with big data. It is oftenly noticed that the enormous volumes of Big Data make sure that the huge amount of information is covered, but analysts cannot just recognize for their valuable content of the data [2]. Predictive data analysis methods found to be erroneous on low order data Collection and issues presented by large data often come in semi-configured or unconfigured and massive. The difference of traditional analytics and big data analytics are shown in below.

	raditional	**Big Data**
Data Sources	It provides structured and static data. It's a homogeneous source of data.	It's a Heterogeneous source of data. It provides unstructured, semi structured and streaming data
Data Storage	Inaccessible proprietary servers are used.	Private / public or hybrid clouds are used.
Database Technology	Relational data stores are used	NoSQL data stores are used
Data Processing	This is Centralized Architecture	This is Distributed Architecture
Analytics	Earliest collected data	real time analytics

Big data analytics can provide new coming to every field. It leads to capable discovery or prediction of new scientific theories, client behavior, social phenomenon, weather guidelines, financial conditions [3] etc., It aids in better conclusion support. Logical area is counting on Big Data analytics.

2. BIG DATA STUDY

Study in big data development for a variety of purposes including successful capture of data, innovative storage solutions and retrieval techniques for this massive data, exploration of cost-effective solutions to move data from storage stage to processing stage**,** Performance Improvement Developed scalable machine learning and data mining algorithms for reduced latency and increased efficiency, accurate learning and accurate predictions, development of new visualization techniques, adherence to strict security and privacy protection strategies and standardization of the big data analysis system. Figure 1 presents an overview of the big data analysis flow. Also, there is a lot of focus on Big Data Analytics in the cloud environment. Big data analysis through social networking is an active part of research to detect sophisticated trends.

Big Data Collection

Big Data is not an example; A sorted human gene is about 140 GB in size [4]; In astronomy, new telescopes generate one PP of data per day [5]. Although Facebook announced an average of 829 billion daily active users in June 2014, it is not difficult to magine the size of the data generated by this public network site. Given the immeasurable resources of Big Data available, we need to focus on reaping its benefits without further delay.

Big Data Storage

With data outburst, Big Data storage systems need to have large and growing capacity, high band width, ability to handle unpredictable load characteristics, reduced I/O path delays, techniques to deal with semi structured and unstructured data without conciliation on consistency and protection. Storage policy can be a federalized one which is simpler and having less communication cost, or a dispersed one which is more consistent and extendable. Although Google File System and Hadoop Distributed File System are so trendy in Big Data circumstances, 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 disintegration may lead to wastage of disk space and creating links for each small file may lead to network delays**.** Load balance in a distributed strategy is an issue that requires consistency in duplicate data. D-copy is very popular for removing unnecessary data for storage capacity upgrade. Since storage drives are slow, to perform analytics on Big Data, in-memory analytics are special as it explore the data is stored in RAM which speeds up the process [7]. Flash [8] introduced the Flexible Distributed Database (RDT), a partition that distributes memory to the hard disk drive, which is firmly embedded in the memory of the cluster nodes to minimize hard disk drive input / output delay. Reuse applications also benefit from stagnation between data that can be used in repetitions. Cloud storage [9] is a popular storage medium for large data, which does not refer to a single medium, but to data stored onsite and accessed online. Many companies are now moving to clouds like Amazon AWS, IBM Smart Cloud and Windows Azure for their big data storage. Cloud provides a hierarchical storage mechanism that uses flash arrays / solid state technology, hard disks and tapes and efficient storage management software, where storage media is selected on a priority basis based on different needs such as delay, cost, energy efficiency, and capacity. And reliability, these basic things are hidden from the end user. Cloud Data Group and Cloud Privacy and Security Protection Issues that make cloud storage intrusive.

Big Database Technology

In [10], Sam Madden discussed how databases can be useful in solving large data problems. Traditional databases, mainly related databases, seem to be the worst choice in scaling, flexibility, fault tolerance and flexibility and distributed systems.

Because Big Data criteria are required to collect and interpret data, data stores, commonly known as NoSQL (not just SQL) systems, are given more importance compared to related databases. NoSQL databases offer flexible planning and flexibility, and can be implemented using cheap materials hardware despite compromises on ACID (nuclear, sustainability, isolation and durability) transactions. Depending on the data model, NoSQL databases can be core value, column-based, document stores, or graph databases. For a key value database, a value is associated with a key, and it has a higher query speed and synchronization compared to the corresponding databases. Dynamotype, Memcache DP, Redis, and Voltmart are examples of key value stores.

OrientDB, Allegro and Virtuso are useful graphical databases for dealing with data that plays a key role in relationships, just like the social database. In the graph, the nodes represent the companies and the edges the relationships.

Databases are similar to key-value stores, but the value or data is stored in documents in some form of markup language like XML. MongoDB and CouchDB. 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], Beth and Tim explain the new database configurations in Big Data, and in [13], models and languages for large data query are discussed, and insights into designing the query engine are provided.

Big Data Processing

MapReduce [14], launched by Google, is the programming model that offers generalization from underlying hardware and make easy parallel programming and execution on multiple clusters. Hadoop [15] is the open source execution of MapReduce, is a popular big data processing engine.. It is initially a block processing system, and it supports performance, reliability and scaling over operational time or delay, introducing some bottlenecks. It also lacks the grip of real-time processing that handles rotten dynamic data [16]. Various improvements to MapRedus' open source implementation have been proposed to adapt it to the needs of large data analyzers and to complete an efficient data center calculation. MapReduce ++ [17] suggests improving response time by calculating the time cost of the task and using a short-work-first planning strategy to execute smaller ones first. Stubby [18] is a workflow optimizer that generates mabrod workflows, which are based on extensibility and transformation, but have the drawback of not considering all sorts of changes.

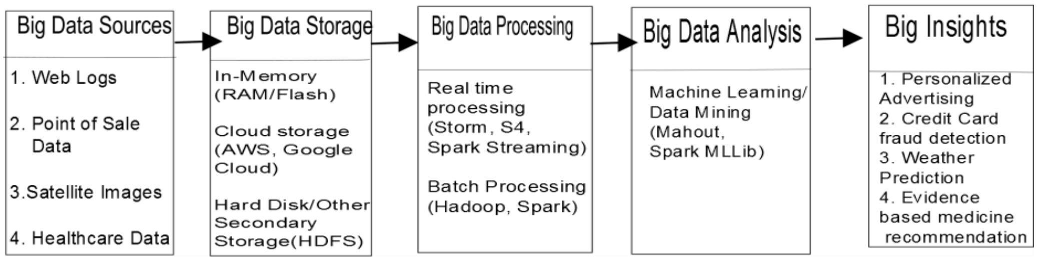
Big Data Sources
1. Web Logs
2. Point of Sale Data
3.Satellite Images
4. Healthcare Data
Big Data Storage
In-Memory (RAM/Flash)
Cloud storage (AWS, Google Cloud)
Hard Disk/Other Secondary Storage(HDFS)
Big Data Processing
Real time processing (Storm, S4, Spark Streaming)
Batch Processing (Hadoop, Spark)
Big Data Analysis
Machine Learning/ Data Mining (Mahout, Spark MLLib)
Big Insights
1. Personalized Advertising
2. Credit Card fraud detection
3. Weather Prediction
4. Evidence based medicine recommendation

Fig 1: Big Data analytic flow

Starfish [19], built on Hoodoop, provides a self-adjusting method for analyzing Big Data, automatically adjusting user requirements and system workloads to give better performance. Radoop [20], Hatoop's integration with Data Miner tool Rapidmin, measures well by increasing data size and combines the benefits of both.

Sailfish [21] uses another abbreviation called Map Reduce Framework, which is 20% to 5 times faster than Hatoop for collecting data and moving data from mapping tasks to minimize tasks.

Twister [22] expands Maybroots by adding reactivation support for Maybrood tasks that facilitate continuous data processing. Holop [23] acts in a similar manner, and has reduced the execution time compared to Hadoop. The impossibility of inducing new data is affected by both limitations such as high response time and volume processing model, which does not apply to real-time stream processing.

Twitter Storm [24] works well for real-time streaming data, where data is delivered to the computer as an infinite package. They are transferred to the consuming processing nodes and the results are published or more data streams are produced for further processing / aggregation. The storm can process millions of data per second. Spark streaming is popularly used to handle streaming data

Big Data Transportation

Although large data analytics can be done effectively in the cloud environment, changing the massive data set to the cloud seems like a challenge. L. Zhang [25] proposed an approach to reduce the online cost of uploading this data to a remote cloud. The paper discusses two online ways to improve data center selection when moving earth-scattered data, and ways to send data to a specific data center.

Big Data Analytics

Data analysis is considerably more difficult compared to data collection and storage. The development of scalable and parallel machine learning methods for online analysis remains a serious challenge [16]. The Apache Mahout program provides several parallel machine learning algorithms tuned for MapReduce implementation, but is often referred to as module processing. In [26], They therefore do not support re-processing or online stream processing, although available integration features can be combined with online processing.

Existing analytics structures are often transaction-based and are effectively used in business applications such as customer segmentation and marketing, financial and accounting operations management, The perspective shifts toward an ecologically based analytics framework that focuses on the integrated analysis of less structured contexts compared to isolated transaction analysis [27].

In [28], the authors present a HACE theorem modeling a large data property and proposing a large data processing model from a data mining perspective. According to the paper, the Big Data conceptual framework consists of three tiers, dealing with Tier 1 data access and computing, Tier II data privacy and domain knowledge and Tier III Big data mining / machine learning algorithms.

Sites built for large-scale data analytics can only handle part of the machine learning algorithms and the best systems support for already established machine learning application events remains an open research question [29]. Traditional data mining algorithms require the entire memory to be loaded into main memory for mining, but moving data to different locations for larger data can be costly. Domain knowledge of applications is also essential as data privacy and data sharing methods may differ depending on the nature and need of the application.

The complex semantic relationship from Big Data improves the performance of applications including search engines and referral systems and provides insights into various social phenomena, but this has become a major challenge due to the diversity and large size of the data.

Security and Privacy Preservation

Protecting the privacy of data in the cloud is a concern because the strategies available to prevent the leakage of sensitive personal information are inadequate. Because companies can benefit from the analysis of important data such as health records and financial transaction records, Failures of traditional privacy protection measures in the cloud can be used to expose data privacy by malicious users, which can severely damage social reputation or cause financial loss to data owners. Data anonymization techniques that hide the identity or other sensitive data of the owners are widely used for privacy protection and many anonymous methods have been proposed. But with big data trends, These anonymous algorithms fail to anonymize such large data sets, and researchers are trying to improve the measurable problems in anonymizing large data sets [30]. A two-stage scalable top-down approach was proposed in [31] to anonymize the large-scale data set for using Maprod on the cloud.

A different privacy paradigm has been in use recently, which protects against the leakage of personal data when processing queries on data. In [32] An algorithm called DFMR has been proposed, which enables the top-Q query of the Mabrut architecture while maintaining a different privacy.

The US government released two reports in May 2014 - the White House report and the PCAST report, mainly dealing with issues related to the privacy of Big Data. These reports deny encryption as the right solution for privacy protection and point out the inadequacies of data anonymization and identification techniques. It provides policy recommendations for responsible and accountable use and disclosure of large data [33]

Visualization and Benchmarking

In the big data era, the development of novel visualization techniques is ongoing and [34] provides a comprehensive review of popular visual analysis systems. [35] Key challenges associated with large data visualization and ways to avoid them are discussed.

Large data visualization methods such as tree maps, circle packing, sunburst, circular network mapping, parallel integrations and stream graphs are analyzed and compared in this study. Described in Topicflow, an interactive visualization tool [36] that must be used in conjunction with Twitter to align and display similar headlines from time to time. Researchers are also concerned about designing a benchmark suite to evaluate the performance of Big Data systems. [37].

Big Data in Cloud Environment

Problems and challenges associated with Big Data processing in the cloud environment are presented in detail from the user, data and hardware perspective [9]. Zimmerman et al. Discusses the service-based organizational classification model for large data in the cloud environment [38]. An analytical model [39] was proposed based on the hierarchical theory to achieve flexibility for the Maybrood jobs running on cloud clusters by giving an estimate of the required resources. Simulation of the sample was performed, but its testing was not validated. In his research paper, Kid [41] proposed a resource management method for using dynamic mabrod clusters in multiple cluster systems, ensuring that results are delivered and planned according to workload characteristics.

Social Network Sourced Big Data

Social networks such as Facebook and Twitter are leading the way in producing big data. Social networking has a strong impact on geophysical technology networks because most of the traffic is contributed by these social networking sites and their affiliates. In [42], The use of social networking analysis has been used to design technical networks and in contrast some methods and cases have been discussed. A user can be a member of multiple networks at once, and these networks form a collective social network where the user can express different behaviors on different networks. At the same time the user can share some common hidden interests across these networks. E. Zhang [43] proposed a proposal for adaptive exchange of information from mixed social networks to predict human behavior that could be used in social marketing, service referrals and customization. Improving the social networking paradigm for gaining knowledge from big data by using participants' personal temporary clouds on social networks to overcome the big data processing challenges [44].

3. SOCIAL ISSUES IN BIG DATA

Companies that now use large data analytics have competitive advantages over those that do not. Cloud-based analytics is very popular due to its scaling and cost effectiveness, but many small companies stay away due to lack of network bandwidth and money.

Companies that now use big data analytics have competitive advantages over non-existent ones. Cloud-based analytics is very popular because of its scaling and cost effectiveness, but many small companies stay away due to lack of network bandwidth and money. Therefore, the same census data can be analyzed and used by the government to determine effective fiscal policies or to use communal sentiments in selecting candidates for the upcoming 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 analysis has proven to be beneficial for large companies, but small companies that previously found it costly and complex are now switching to use its merits [46].

4. CONCLUSION AND FUTURE SCOPE

Research on various aspects of Big Data accelerates the capture of Big Data flow rates. While this article gives some flavor of Big Data research progress, many research questions in this area are still open to future research assignments. There is a growing need to address issues such as data and tool dynamics for different data formats and tools and the integration of various large data analysis structures. Protecting the privacy and privacy of sensitive information, especially cloud computing, is another area of great concern to researchers. Data acquisition requires an efficient real-time analysis framework for the future of streaming. Real-time security monitoring is a challenging task that requires more research attention. Further improvements are needed to define appropriate programming summaries for domain-specific applications and to develop optimal symmetric planning strategies for large-scale data sets. The research aims to develop scalable and parallel machine learning methods for in-depth learning and complex large data that attract more attention. Further research should be aimed at finding effective and practical solutions to these problems. It is certain that valuable treasures of knowledge hidden under the deep layers of the Big Data Ocean will continue to attract researchers for many years to come.

4. 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.