

BIG DATA ANALYTICS AND DECISION MAKING: TECHNIQUES, TECHNOLOGIES AND CHALLENGES

Abstract

The age of big data has come with voluminous, complex, noise and ever fast growing data from technology driven multiple data sources. Big data encompasses the domains such as healthcare, biomedical, government, research and commerce. Researchers from diverse fields have been focusing on the big data to obtain quality knowledge of contribution in their field. The researchers has reported to use many modern tools and techniques specific to big data in solving the big data decision making task. This paper is focused at (i) concepts and features associated to big data, (ii) the state-of-the-art techniques and tools for big data decision making, and (iii) to discover the challenges pertaining to big data so that future researchers can obtain the directions from it.

Keywords: Big data, Decision making; Data analysis; data-intensive processing;

Authors

Tanvir Habib Sardar

Assistant Professor
Department of Computer Science and Engineering
GST, GITAM University
Bangalore, India
tsardar@gitam.edu

Rashel Sarkar

Department of Computer Science & Engineering
University of Science and Technology
Meghalaya, India
sarkarkarrashel@gmail.com

Sk Jamil Ahmed

Department of Computer Science & Engineering
Alvas Institute of Engineering and Technology
Moodbidri, Karnataka
jamil.pace@gmail.com

I. INTRODUCTION

The exponential growth of data has arrived from the day-to-day human intractable or passively assistive tools such as sensors, scientific laboratories, e-commerce or government firms, and social media websites, etc. The growth of this data has already surpasses the capabilities of microprocessor based processing as the microprocessor's growth is slowed down far below to the Moore's law. A decade below, in 2011, 2.5 quintillion data bytes were created [1]. It is observed by IBM that 90% of world data is created in each 2 years and exceeding. Six thousand Walmart stores produces above 267 million of online transactions. In a single day a survey Telescope's image data consumes up to 30 trillion bytes [2]. Alibaba produces 20 terabytes of data by over 880 million online transactions. Figure 1 shows the data volume chart as predicted by International data corporation (IDC) [3]. These data brings us to the definitive conclusion that the era of big data has arrived.

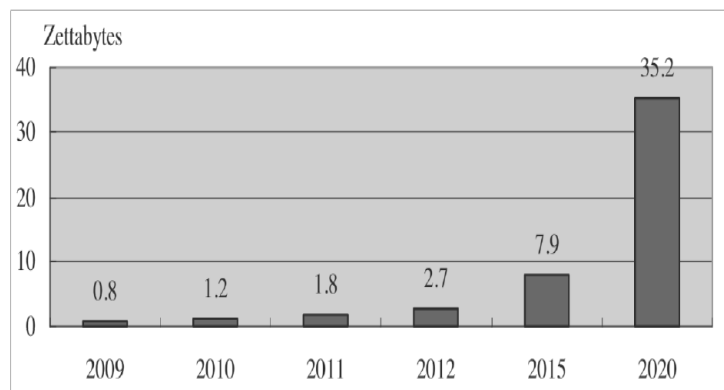


Figure 1: Global Data Volume Prediction by IDC

Along with the voluminous size of the datasets, big data also defined by the complex data structures and by a challenging technique of obtaining and managing data [4]. These challenges associated with big data is aimed at dealing effectively by scientist and engineers, known as Big Data Scientists and Big Data Engineers respectively. Followed and motivated by the Obama administration in 2012, Gartner put on the big data among the top ten technology trends and critical technology trends.

The analysis of big data can derive into a big value chain for human life and business profit through the process of big data identification, integration, and exploitation [5]. The researchers in the big data domain has consistently developing tools and proposing technological advancement to process the big data and obtain perfect decision making. The decision making in big data domain has already been experimented successfully in many domains such as business sale analysis, development of customer specific products, loyalty, patient discovery and clinical decision making, tourism marketing, and transportation etc. Some survey reported that over half of the 560 surveyed institutions and enterprises have intuition and understanding that big data decision making would help them in enhancing the efficiency of operation, obtaining the strategic decision, better customer interaction, advertisement and retention, and setting the product price etc. [6] [2].

Decision support is the key strategy to be included in every sources of big data. The theory of decision in general means as discovering values, logics, rationalities, probabilistic model deriving and other relative challenges in the field of economics and computer science using optimal strategy and techniques of mathematics and statistical modelling. Normative theory of decision making refers to the optimal usage of techniques, technologies and tools to discover the best decision in a framework of rational bound. Under this definition, the decision making in big data domain belongs to each steps of big data accumulation to prediction of classified knowledge system, including the visualization process.

There are many technique and tools developed solely for decision making from big data, although sometimes they fail to provide optimal solutions. For examples, some of these decision making tools involved with multi-disciplined areas such as optimization methods, statistics, machine and deep learning, visualization, data mining, and social media analysis etc. Along with this the big data tools can be divided into three types such as batch processing, stream processing and hybrid processing tools [2]. Figure 2 explains the relationship between data science and decision support system (DSS).

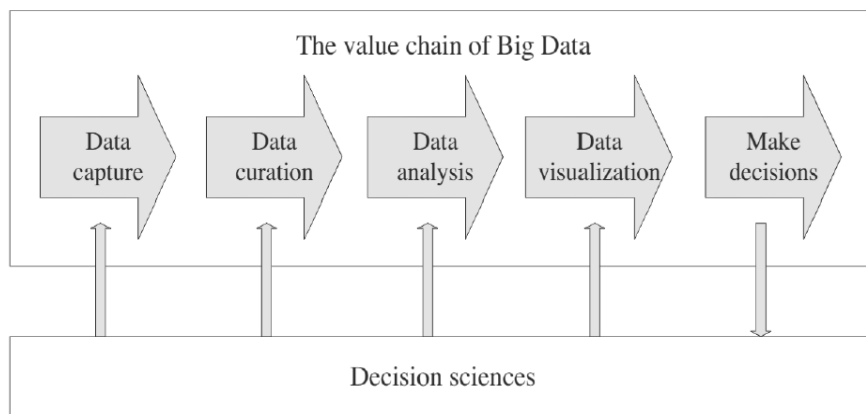


Figure 2: Relationship between Data Science and DSS

The paper is organized as below. The section 2 explains the techniques and tools being used in the big data processing. The section 3 explains challenges pertaining to big data to get the future direction in research. The last section concludes the work.

II. TECHNIQUES AND TOOLS OF BIG DATA PROCESSING

A wide variety of tools are developed by scientists and big data enterprises to obtain the valuable knowledge for their decision making need fulfillment. The tools has crossed multidiscipline domains and proven effective for discovering, capturing, integrating, analysis and visualizing the big data. This section provide a comprehensive review of some current trends and development in big data decision making techniques and tools. The paradigm for big data decision making is shown in figure 3.

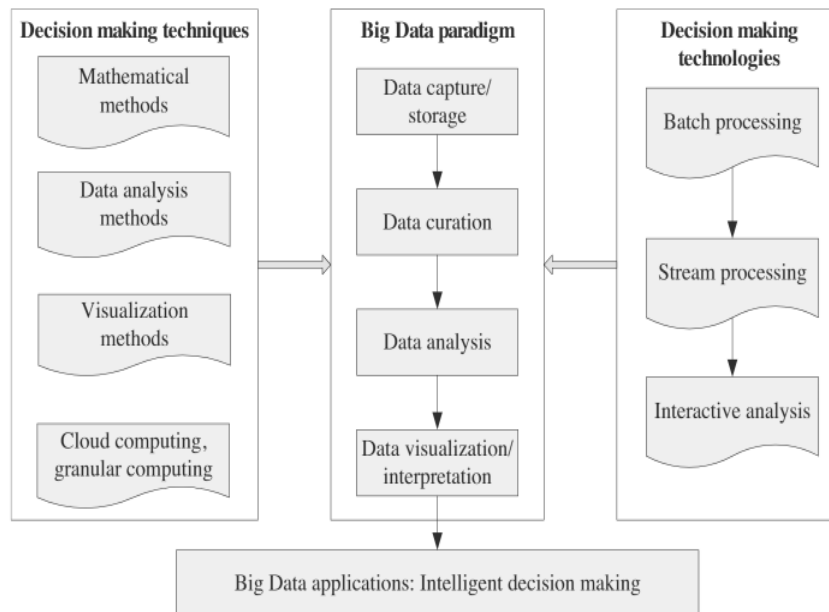


Figure 3: Paradigms of Big Data Processing

1. Techniques of big data decision making: The techniques of decision making in big data field is either developed or in developing stage due to enormous complexing in each phase of big data processing. These techniques are usually picked up based on the specific applications and their respective objectives. There are many techniques which involves many disciplines and generally overlap with each other. The categories are explained below one by one.

- **Mathematical techniques:** The big data study has taken two major concepts from the field of mathematics: statistics and optimization techniques. The data gathering and analysis or raw data is mainly gone through the statistical approaches before the actual process of data analysis starts. The numerical description of data points and to derive the casual relationships or co-relationships between objectives are derived using statistical measures. To deal with huge volume of big data, some modifications of big data specific statistical techniques are developed such as parallel statistics [7], statistical computing [8], and statistical learning [9].

Due to effectiveness in quantitative, the optimization techniques is being used in big data. The issues of high memory cost and time consumption, the optimization techniques are generally incorporated with data reduction techniques [10] or parallelism [11]. Several studies of co-evolutionary algorithms [12], real-time optimization [13] in wireless sensor networks [14] and intelligent transportation systems [15] has shown remarkable use of optimization techniques in big data decision making.

- **Data analysis techniques:** Data analysis takes the advantages of different beneficial techniques for big data, developed under different domains such as data mining, machine learning, artificial neural networks, and signal processing etc.

Data mining is a set of techniques offered to obtain useful patterns and information from the datasets. Big data mining requires to improve the existing data mining techniques. For this, generally parallel implementations [16] of data mining algorithms and dimensionality reduction methods [17] are also considered. The most widely used data mining techniques used in big data are clustering and regressing. However, classification and fuzzy logic based data modelling is also reported to be used in the literature.

Machine learning is a sub field of artificial intelligence. Machine learning is classified into supervised learning and unsupervised learning. In supervised learning the output data labels are known for a certain input set. A well designed supervised machine learning algorithm learns the big data inputs for the specific output assignment, as a training part. After the training process, the algorithm can automatically classify the big data inputs to its respective output class. In unsupervised learning, the data are not labelled. The unsupervised learning algorithms automatically classify the input big data into different groups based on inherent similar features in the dataset. The machine learning algorithms are applied successfully in many big data types such as biological big data [18], sensors data [19], and stock market [19] for example.

Artificial neural network is widely used method of learning useful patterns from the datasets using distributed node based processing which emulates the human brains. The big data processing using artificial neural network is however a challenging task due to the artificial neural network's requirement of many hidden layers and nodes construction for a well performed learning mechanism. However, more hidden layers and nodes consumes more time and memory of computing model which has a deep adversarial effect to the big data analysis. This adversarial effect is however addressed by proposing (i) a sampling method which reduces the sizes of the big data before input to the artificial neural network and (ii) by designing a parallel or distributed artificial neural network for big data processing [20]. To contain big data learning, many specific deep learning methods are also designed by the researchers [21]. It is interesting to know that many deep learning models has successfully dealt for different big data applications such as drug discovery [22], genomic medicine [23], and text mining [24].

Textual big data, especially from social media and ecommerce websites, is widely used for analysis of human sentiments. This process of learning involved few machine learning and lexicon-based methods and known as sentiment analysis [25]. The sentiment analysis is aimed at different applications such as subjectivity classification [26], polarity determination [27], spam detection [28], review usefulness measurement [29], aspect extraction [30], and so on. To implement sentiment analysis for the big data, some new big data platforms are chosen such as MapReduce and Storm [31].

- **Visualization techniques:** The visualization technique refers to the ways required for displaying the information using intuitive display like tables, images, and diagrams. Every big and small firms uses the visualization to display the compact but meaningful explanation of the datasets and knowledge. For example, Facebook uses

timeline as a visualization method for manipulate and display its data in its dataset. Big data visualization would increase the meaningfulness but the complexity of big data make it a difficult cases than the traditional datasets [32]. To deal with big data, many modern visualization techniques tries to discover proper visualization after reducing the dataset size by feature extraction and parallel way of execution [33] [34]. The best visualization is mandatory for big data as a good visualization demonstrate a good visualization is better than a thousand of petabytes [35].

- **Cloud computing:** Cloud computing is a new revolution in the computing where the required resources and service is rented to the end users, without requiring to actually buy the resources or the software and installing these to the user end. The cloud computing model is suitable for big data processing. The decision support can be accomplished by cloud computing is the data management, tuning of models, data quality, and data currency is properly maintained [36].
 - **Fuzzy sets and systems:** Fuzzy set is an effective solution for many big data problems. Fuzzy set and logic is good for dealing with uncertainties and vagueness in the dataset, which is very common for big datasets [37]. The fuzzy logic is thus made to be a good choice for extracting knowledge from incomplete big data [38]. Fuzzy systems are even used for deep learning in big data analysis and applications as well [39]. The fuzzy techniques being used for big data are evolving fuzzy systems [40], neural fuzzy classifier [41], linguistic fuzzy rule-based classifier [42], and fuzzy C-Means is proven best for clustering the big data [43]. Also, many pattern recognition algorithm such as fuzzy inference systems, fuzzy Bayesian process [44], and fuzzy query system [45] are also experimented for big data applications. Big data dimensionality reduction can also be performed neural fuzzy classifiers [46].
2. **Technologies of big data decision making:** The characteristics and features of big data requires novel technologies, in terms of infrastructures and platforms, for providing timely and accurate decision making. Figure 4 shows a historical perspective of the big data framework of technologies. MapReduce changed the batch processing framework and brought a revolution in batch based big data processing [47], proposed by google in 2003. However, that time only large datasets were being processed rather than the big data.

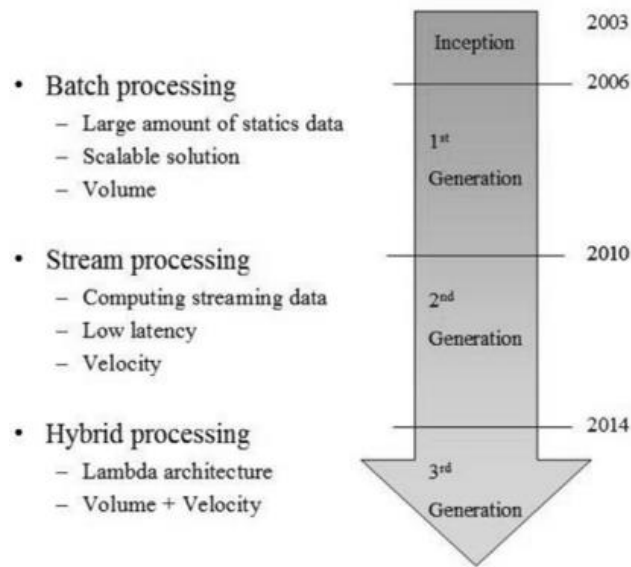


Figure 4: The Three Generations of Processing Paradigms

With the arrival of Hadoop in 2006, the first generation of big data processing started. Hadoop uses MapReduce as its processing engine. The second generation of big data processing was started by S4 (a Yahoo product of 2010). S4 dealt with both the static and big data. The hybrid processing can bring us to the third generation. However, the enough development is this area is yet to happen to let us inter into the third generation. Table 1 simplifies with its visualization regarding the three generations of technologies by the detailed processing technologies.

Table 1: Three Generations of Big Data Technologies

Paradigm	Technology
Batch Processing	MapReduce
	Hadoop
	Flume
	Scribe
	Dryad
	Apache Mahout
	Jaspersoft BI Suite
	Pentaho
	Skytree Server
	Cascading
	Spark
	Tableau
	Karmasphere
	Pig
Sqoop	
Stream Processing	Kafka
	Flume

	Kestrel
	Strom
	S4
	SQLstream
	Splunk
	SAP Hana
	Spark Streaming
Hybrid Processing	Lambdoop
	SummingBird

Batch processing takes care of the data which is stored in storage. The advantages associated with batch processing are scalability and reliability. The scalability is achieved by parallel implementations like that of MapReduce. The stream processing on the other hand processes big data in real time. This paradigm takes a diskless processing approach to achieve low latency. The hybrid processing synthesizes both the batch and stream processing based on Lambda architecture [48].

III. CHALLENGES PERTAINING TO BIG DATA

The ultimate target is to develop the big data solutions for decision making which were never before available. In this section, we shall discuss the challenges in big data decision making and the future solutions to it. There are many factors which influence the decision making process for big data. The literature reports and studies show that the factors and their impacts change over time. The big data and its analysis for decision making is an ad-hoc process where the organizations' changes are frequently altered for obtaining quality output. The agreements are changed to obtain big data, new staff are hired and new departments are formed so as to obtain advantages by discovering features from big data and subsequent decision making in a short span of time. The factors which affect the decision making from big data are listed in table 2.

Table 2: Factors Influencing the Decision Making Quality

Factors	Description
Contractual governance	The big data's quality can be enhanced if the proper agreement can be laid down with the data source provider. Agreements guarantee collaborative sharing of data with proper understanding, by defining responsibilities and methods, through a good communication.
Relational governance	Relational governance is mandatory for developing trust between enterprise stakeholders and for guaranteeing the sharing of required knowledge that is mandatory to understand big data. Good relational governance comprises communication and knowledge alteration which is essential to comprehend and process data.
Big data analytics capabilities	Analyzing big data can cover loads of variables and constraints. It was hard to discover the correct tools. Which methods can be probably used and in what way big data can be pictured is a test. This was frequently an extensive exploration procedure in which knowledge of big data and the domain was needed.

Knowledge exchange	Together data and knowledge around the data desired to be conveyed. Knowledge around in what way the data is composed and handled is essential for being able to understand the data and to comprehend by what means it can be used. When big data analyst hold more knowledge regarding the background, the use and the discovery of patterns and associations would become easier.
Collaboration and Cooperation	The capability to cooperate amongst big data benefactors, analysts and decision-makers is a main need to create a big data chain. Besides, the incapability to cooperate with data benefactors can reduce the paths which is used for generating valuable knowledge.
Process integration and standardization	The capacity to assimilate procedures and to normalize responsibilities and outputs in improving the big data chain. This consequences in lesser labors and price to usage big data analytics.
Routinizing and standardization	By routinely updating the big data collection from sources would improve the big data velocity matching. This is mandatory for big data analysis in real time.
Flexible infrastructure	A flexible infrastructure increases the capacity to handle the big data for subsequent processing.
Staff	A specialist and expert staff in the field of big data and analytics has enough knowledge for data gathering related communication making and interpreting the big data analytics results.
Data quality of the big data sources	The data quality is very important. An error data can generate costly wrong decisions. These wrong decisions becomes very bad consequence maker for critical sectors.
Decision-maker quality	Decision-makers should interpret the result of the big data analytics and comprehend the meaning. The experienced decision-makers are the faster decisions makers.

The main challenges discovered in decision making can relate to the velocity, validity and veracity and these are connected with the following:

1. **Processing:** The velocity of data sometimes let the application obtain and deal with just a part of data, leaving another part behind. This makes the decision making poor as the entire picture of the dataset becomes not so clear. For example, some part of the data which shows a behavior like fraudulent becomes unknown if that part of data in unavailable.
2. **Noise:** The presence of noise creates a problems on data perception and to obtain key insight becomes a problem in case of noise presence.
3. **Error:** In many cases the source only has the information on the context of the data. The data analytics have no idea on data context in this cases. For example, a data may be collected two years back and it reflect the scenarios of last two years but wrongly it was communicated that the data is of last years. This is an error of data context and the decision making from this data would be wrong.

IV. CONCLUSION

Big data is a popular domain which is still developing in an enormous speed. The big data decision making is a new sub-domain of big data analysis which encompasses the techniques and technologies of many other domains. The techniques and technologies of big data is presented in this paper while providing textual details and also using few figures and tables also. The paper then provides the challenges associated with big data processing and discovers factors which influences the big data decision making from the literature.

REFERENCES

- [1] Hilbert, Martin, and Priscila López. "The world's technological capacity to store, communicate, and compute information." *science* 332, no. 6025 (2011): 60-65.
- [2] Chen, CL Philip, and Chun-Yang Zhang. "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data." *Information sciences* 275 (2014): 314-347.
- [3] Tien, James M. "Big data: Unleashing information." *Journal of Systems Science and Systems Engineering* 22, no. 2 (2013): 127-151.
- [4] R. Casado *et al.* Emerging trends and technologies in big data processing, *Concurr. Comp-Pract. E.*, (2015).
- [5] Miller, H. Gilbert, and Peter Mork. "From data to decisions: a value chain for big data." *It Professional* 15, no. 1 (2013): 57-59.
- [6] Manyika, James, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung Byers. *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute, 2011.
- [7] Mahani, Alireza S., and Mansour TA Sharabiani. "SIMD parallel MCMC sampling with applications for big-data Bayesian analytics." *Computational Statistics & Data Analysis* 88 (2015): 75-99.
- [8] Wilkinson, Leland. "The future of statistical computing." *Technometrics* 50, no. 4 (2008): 418-435.
- [9] Hastie, Trevor, Robert Tibshirani, Jerome H. Friedman, and Jerome H. Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. New York: springer, 2009.
- [10] Yan, Jun, Ning Liu, Shuicheng Yan, Qiang Yang, Weiguo Fan, Wei Wei, and Zheng Chen. "Trace-oriented feature analysis for large-scale text data dimension reduction." *IEEE Transactions on Knowledge and Data Engineering* 23, no. 7 (2010): 1103-1117.
- [11] Sahimi, Muhammad, and Hossein Hamzehpour. "Efficient computational strategies for solving global optimization problems." *Computing in Science & Engineering* 12, no. 04 (2010): 74-83.
- [12] Li, Xiaodong, and Xin Yao. "Cooperatively coevolving particle swarms for large scale optimization." *IEEE Transactions on Evolutionary Computation* 16, no. 2 (2011): 210-224.
- [13] Sardar, Tanvir Habib, and Zahid Ansari. "An analysis of MapReduce efficiency in document clustering using parallel K-means algorithm." *Future Computing and Informatics Journal* 3, no. 2 (2018): 200-209.
- [14] Sardar, Tanvir Habib, and Zahid Ansari. "Partition based clustering of large datasets using MapReduce framework: An analysis of recent themes and directions." *Future Computing and Informatics Journal* 3, no. 2 (2018): 247-261.
- [15] Sardar, Tanvir Habib, and Zahid Ansari. "Detection and confirmation of web robot requests for cleaning the voluminous web log data." In *2014 International Conference on the IMPact of E-Technology on US (IMPETUS)*, pp. 13-19. IEEE, 2014.
- [16] Sardar, Tanvir Habib, and Zahid Ansari. "An analysis of distributed document clustering using MapReduce based K-means algorithm." *Journal of The Institution of Engineers (India): Series B* 101, no. 6 (2020): 641-650.

- [17] Ansari, Zahid, Asif Afzal, and Tanvir Habib Sardar. "Data categorization using hadoop MapReduce-based parallel K-means clustering." *Journal of The Institution of Engineers (India): Series B* 100, no. 2 (2019): 95-103.
- [18] Wen, Zhenshu, Wanwei Zhang, Tao Zeng, and Luonan Chen. "MCentridFS: a tool for identifying module biomarkers for multi-phenotypes from high-throughput data." *Molecular BioSystems* 10, no. 11 (2014): 2870-2875.
- [19] Wang, Yi, Xinli Jiang, Rongyu Cao, and Xiyang Wang. "Robust indoor human activity recognition using wireless signals." *Sensors* 15, no. 7 (2015): 17195-17208.
- [20] Nedjah, Nadia, Felipe P. da Silva, Alan O. de Sá, Luiza M. Mourelle, and Diana A. Bonilla. "A massively parallel pipelined reconfigurable design for M-PLN based neural networks for efficient image classification." *Neurocomputing* 183 (2016): 39-55.
- [21] Arel, Itamar, Derek C. Rose, and Thomas P. Karnowski. "Deep machine learning-a new frontier in artificial intelligence research [research frontier]." *IEEE computational intelligence magazine* 5, no. 4 (2010): 13-18.
- [22] Preuer, Kristina, Richard PI Lewis, Sepp Hochreiter, Andreas Bender, Krishna C. Bulusu, and Günter Klambauer. "DeepSynergy: predicting anti-cancer drug synergy with Deep Learning." *Bioinformatics* 34, no. 9 (2018): 1538-1546.
- [23] Sardar, Tanvir Habib, Ahmed Rimaz Faizabadi, and Zahid Ansari. "An evaluation of MapReduce framework in cluster analysis." In *2017 International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICT)*, pp. 110-114. IEEE, 2017.
- [24] Sardar, Tanvir Habib, Zahid Ansari, and Amina Khatun. "An evaluation of Hadoop cluster efficiency in document clustering using parallel K-means." In *2017 IEEE International Conference on Circuits and Systems (ICCS)*, pp. 17-20. IEEE, 2017.
- [25] Sardar, Tanvir Habib, Amina Khatun, and Sahanowaj Khan. "Design of energy aware collection tree protocol in wireless sensor network." In *2017 IEEE International Conference on Circuits and Systems (ICCS)*, pp. 12-17. IEEE, 2017.
- [26] Ansari, Zahid, Tanvir Habib Sardar, Moksud Alam Mallik, and Naveen D. Chandavarkar. "Data mining in soft computing framework: a survey." (2002).
- [27] Sardar, Tanvir Habib, and Ahmed Rimaz Faizabadi. "Parallelization and analysis of selected numerical algorithms using OpenMP and Pluto on symmetric multiprocessing machine." *Data Technologies and Applications* (2019).
- [28] Siddiqa, Noor, and Tanvir Habib Sardar. "Multi-Layered Security System Using Cryptography and Steganography." (2019).
- [29] Sardar, Tanvir Habib, Zahid Ansari, Naveen D. Chandavarkar, and Amjad Khan. "A Methodology for Detecting Web Robot Requests."
- [30] Sardar, Tanvir Habib. "A Methodology in Mobile Networks for Global Roaming." *Oriental Journal of Computer Science and Technology* 6, no. 4 (2013): 391-396.
- [31] Sardar, T. Habib, Zahid Ansari, and Amjad Khan. "A Methodology for Wireless Intrusion Detection System." *International Journal of Computer Applications* 975: 8887.
- [32] Staff, C. A. C. M. "Visualizations make big data meaningful." (2014): 19-21.
- [33] Bennett, Janine Camille, David Thompson, Joshua Levine, Peer-Timo Bremer, Attila Gyulassy, Valerio Pascucci, and Philippe Pierre Pebay. *Analysis of Large-Scale Scalar Data Using Hixels*. No. SAND2011-8450C. Sandia National Lab.(SNL-CA), Livermore, CA (United States), 2011.
- [34] Ahrens, James, Kristi Brislawn, Ken Martin, Berk Geveci, C. Charles Law, and Michael Papka. "Large-scale data visualization using parallel data streaming." *IEEE Computer graphics and Applications* 21, no. 4 (2001): 34-41.
- [35] Childs, Hank, Berk Geveci, Will Schroeder, Jeremy Meredith, Kenneth Moreland, Christopher Sewell, Torsten Kuhlen, and E. Wes Bethel. "Research challenges for visualization software." *Computer* 46, no. 5 (2013): 34-42.
- [36] Assunção, Marcos D., Rodrigo N. Calheiros, Silvia Bianchi, Marco AS Netto, and Rajkumar Buyya. "Big Data computing and clouds: Trends and future directions." *Journal of parallel and distributed computing* 79 (2015): 3-15.

- [37] Morente-Molinera, Juan Antonio, Ignacio J. Pérez, M. Raquel Ureña, and Enrique Herrera-Viedma. "Creating knowledge databases for storing and sharing people knowledge automatically using group decision making and fuzzy ontologies." *Information Sciences* 328 (2016): 418-434.
- [38] Lin, Chun-Wei, and Tzung-Pei Hong. "A survey of fuzzy web mining." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 3, no. 3 (2013): 190-199.
- [39] Kaburlasos, Vassilis G., and George A. Papakostas. "Learning distributions of image features by interactive fuzzy lattice reasoning in pattern recognition applications." *IEEE Computational Intelligence Magazine* 10, no. 3 (2015): 42-51.
- [40] Iglesias, José Antonio, Alexandra Tiemblo, Agapito Ledezma, and Araceli Sanchis. "Web news mining in an evolving framework." *Information Fusion* 28 (2016): 90-98.
- [41] Chang, Hsien-Tsung, Nilamadhab Mishra, and Chung-Chih Lin. "IoT big-data centred knowledge granule analytic and cluster framework for BI applications: a case base analysis." *PloS one* 10, no. 11 (2015): e0141980.
- [42] López, Victoria, Sara Del Río, José Manuel Benítez, and Francisco Herrera. "Cost-sensitive linguistic fuzzy rule based classification systems under the MapReduce framework for imbalanced big data." *Fuzzy Sets and Systems* 258 (2015): 5-38.
- [43] Lu, Hua-pu, Zhi-yuan Sun, and Wen-cong Qu. "Big data-driven based real-time traffic flow state identification and prediction." *Discrete Dynamics in Nature and Society* 2015 (2015).
- [44] Ramachandramurthy, Sivaraman, Srinivasan Subramaniam, and Chandrasekeran Ramasamy. "Distilling big data: refining quality information in the era of yottabytes." *The Scientific World Journal* 2015 (2015).
- [45] Wang, Hai, Zeshui Xu, Hamido Fujita, and Shousheng Liu. "Towards felicitous decision making: An overview on challenges and trends of Big Data." *Information Sciences* 367 (2016): 747-765.
- [46] Azar, Ahmad Taher, and Aboul Ella Hassanien. "Dimensionality reduction of medical big data using neural-fuzzy classifier." *Soft computing* 19, no. 4 (2015): 1115-1127.
- [47] Dean, Jeffrey, and Sanjay Ghemawat. "MapReduce: Simplified data processing on large clusters." (2004).
- [48] Janssen, Marijn, Haiko van der Voort, and Agung Wahyudi. "Factors influencing big data decision-making quality." *Journal of business research* 70 (2017): 338-345.