International Journal of Electrical and Computer Engineering (IJECE)
Vol. 8, No. 3, June 2018, pp. 1854~1862
ISSN: 2088-8708, DOI: 10.11591/ijece.v8i3.pp1854-1862

A Survey of Machine Learning Techniques for Self-tuning Hadoop Performance

Md. Armanur Rahman[1], J. Hossen[2], Venkataseshaiah C[3], CK Ho[4], Tan Kim Geok[5], Aziza Sultana[6], Jesmeen M. Z. H.[7], Ferdous Hossain[8]
[1,2,3,5,7,8]Faculty of Engineering and Technology, Multimedia University, Melaka, 75450, Malaysia
[4]Faculty of Computing and Informatics, Multimedia University, Cyberjaya, 63100, Malaysia
[6]Faculty of Computing and Engineering, Dhaka International University, Dhaka, 1205, Bangladesh

Article Info

Article history:

Received Jan 29, 2018
Revised Mar 20, 2018
Accepted Mar 30, 2018

Keyword:

Hadoop
HDFS
Machine learning
MapReduce
Parameter

ABSTRACT

The Apache Hadoop framework is an open source implementation of MapReduce for processing and storing big data. However, to get the best performance from this is a big challenge because of its large number configuration parameters. In this paper, the concept of critical issues of Hadoop system, big data and machine learning have been highlighted and an analysis of some machine learning techniques applied so far, for improving the Hadoop performance is presented. Then, a promising machine learning technique using deep learning algorithm is proposed for Hadoop system performance improvement.

Corresponding Author:

Md. Armanur Rahman,
Faculty of Engineering and Technology,
Multimedia University,
Melaka, 75450, Malaysia.
Email: arman.bdmail@gmail.com

1. INTRODUCTION

The purpose of this paper is threefold:
To provides a brief description of the concept of machine learning, big data and Hadoop system. To present a systematic analysis of existing techniques in terms of performance, parameters, dataset and system configuration. To propose a promising technique using deep learning algorithm for improving the Hadoop system performance in processing big data.

A roadmap of this paper is given in Figure 1. In Section 1 and Section 2, Hadoop system with MapReduce (MR), Hadoop Distributed File System (HDFS) and YARN have been discussed. Then, a discussion about big data and V's, classification of machine learning and existing machine learning algorithms is presented. In Section 2.2, critical issues in Hadoop system are discussed. In Section 3, a promising application of Deep Learning algorithm to improve the Hadoop performance is discussed. A summary is presented in Section 4.

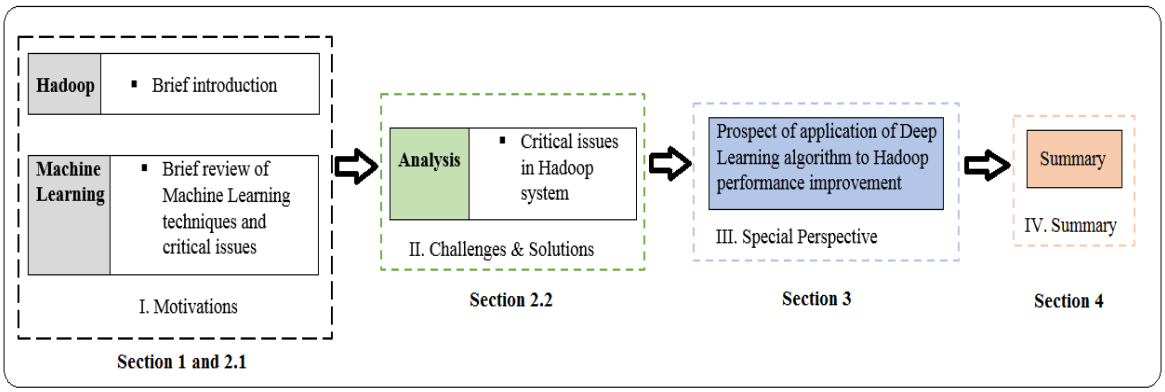

Figure 1. Roadmap of this survey

1.1 Hadoop system

The Apache Hadoop framework mainly consists of three component: MR, HDFS and YARN. The role of MR is data processing. The role of HDFS is to manage storage which is done by breaking a file into multiple blocks and copying each of them into three different servers [1]-[4]. MR is considered as a programming process that comprises of JobTracker (manage the task) and TaskTracker (run the task).

Figure 2 shows the high-level architecture of Hadoop. Hadoop works with two nodes namely master node and slave node. Under the master node, there are TaskTracker and JobTracker in MR Layer and NameNode and DataNode in HDFS layer. Under slave node, there are TaskTracker in MR layer and DataNode in HDFS layer. TaskTracker of Master Node contacts with the JobTracker and JobTracker contacts with the slave node TaskTracker. NameNode contacts with all DataNode.

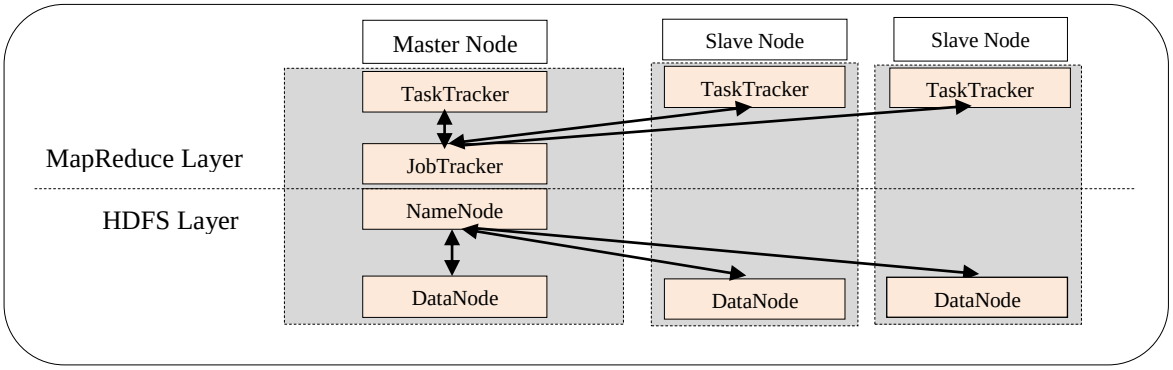

Figure 2. High level Architecture of Hadoop

1.1.1. Hadoop file system (HDFS)

HDFS is designed for reliable and efficient storage of very large datasets [5]. Hadoop enables scaling to a large number of hosts and data partitioning in many nodes for performing computations. Storage of file system metadata and application data are done by separately in HDFS [3]. HDFS stores metadata in a dedicated server called Name node. Application data is stored on another server called Data Node [6], [7].

1.1.2. MapReduce

MR is a programming process which contains two functions namely map and reduce [6], [8]. In terms of designing, MR comprises of three main components: programming, storage designing, and scheduling. In programming step, map function receives input as value and key from user and passes the output to reduce function for further processing and generating the result [8], [3]. The reduce function processes the output from the map function through shuffling, sorting and merging of data [7], [9]. Figure 3 shows MR work process.

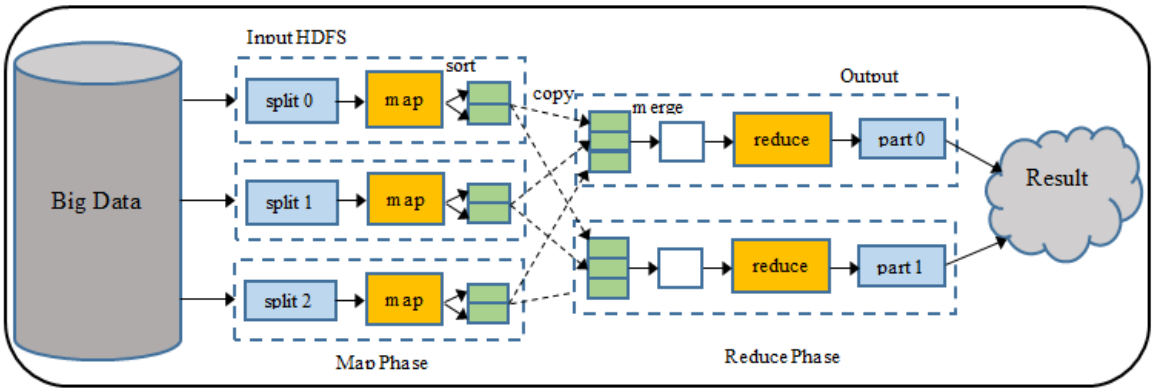

Figure 3. MapReduce process

1.1.3. Parameter

Hadoop contains over 200 parameters where the settings of these parameters determine it effective performance [10-15]. Among these, about 30 parameters can directly impact the performance.

2. MACHINE LEARNING TECHNIQUES AND CRITICAL ISSUES

2.1. Classification of machine learning algorithms

Machine Learning (ML) simply refers to the intelligence of a machine where the machine can provide decision [16], [17]. It has greatly impacted information science in the sectors of prediction, classification, image recognition, computer vision, speech processing, natural language understanding, neuroscience, health, and IoT (Internet of Things). ML algorithm is required to process information from the verity of data within a limited time duration. It is challenged by the emergence of big data [18], [7]. Machine Learning is categorized into supervised, unsupervised, semi-supervised and reinforcement learning [19]. Most popular machine learning algorithms are shown in Figure 4.

a. Supervised Learning: Supervised learning is skilled by labelled instances, like an input as the expected result is known. Supervised learning delivers dataset comprising of both structure and labels.
b. Unsupervised Learning: Unsupervised learning conducted data where no previous labels and its aim is to discover data and trace similarities among the objects. This is a technique of exploring labels since the data itself. Unsupervised learning functions well on the transactional dataset.
c. Semi-supervised Learning: Semi-supervised learning and supervised learning can use the same application but semi-supervised learning can do together with labeled to unlabeled data because of learning.
d. Reinforcement Learning: Reinforcement learning is frequently used in navigation, gaming and robotics. This learning method which connects by a dynamic situation in where it has to perform a particular aim except a trainer explicitly saying it whether has approached its aim [20].

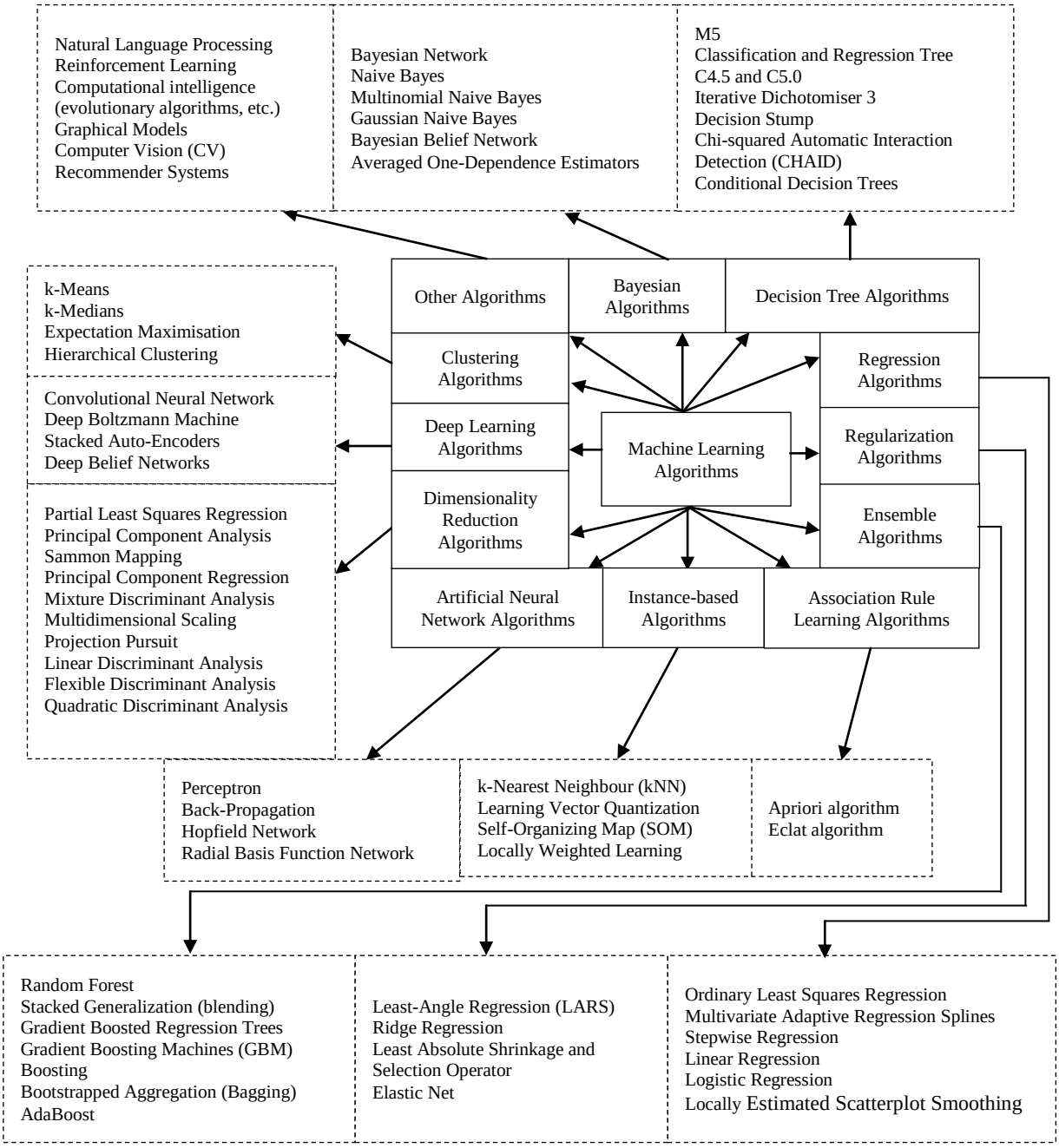

Figure 4. Machine Learning Algorithms

2.2. Analysis of critical issues in hadoop system

The critical issues in Hadoop system for big data processing are depicted in Figure 5.

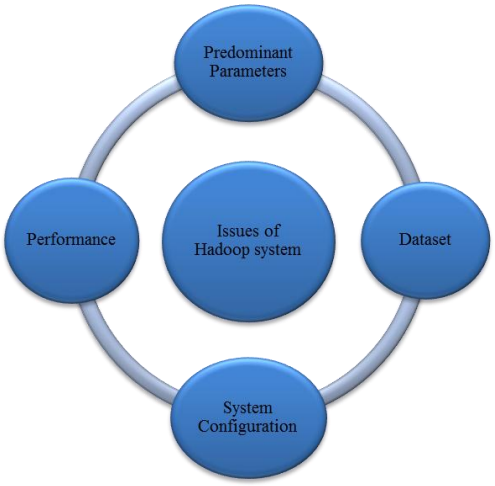

Figure 5. Issues of Hadoop system for Big data processing

2.2.1. Performance

A number of models have been published in the last few years on Hadoop and MR system to optimize the performance by different types of self-tuning techniques. Comparison with default Hadoop systems is made to find out the parameters involved in managing system performances, understand the way those parameters work and look for the limitations of default Hadoop systems. In order to improve Hadoop system performance for processing big data, an efficient algorithm is crucial for optimizing the performance. Many researchers have developed a different algorithm for improving Hadoop system performance compared to the default configuration.

Different type of ML algorithm have been applied in Hadoop system for performance improvement. Algorithms are:

a. Random-Forest
   One of the popular algorithms of ML is Random Forest. Among many other approaches for Hadoop performance improvement, Random-Forest Approach to Auto-Tuning Hadoop's Configuration (RFHOC), is a model to tune the configuration parameters automatically in optimizing the performance for a specific application that runs on a particular cluster. The RFHOC establishes two models based on the random-forest approach that work with the map and reduce stages in a similarly. Five Hadoop programs namely wordcount, terasort, sort, Adjlist and Inverted-Index are used in the evaluation of RFHOC. The evaluation shows that the performance has been speeded up by an average factor of 2.11times where the maximum speed can run up to 7.4 times compared to cost-based optimization (CBO) approach [1].
b. Support Vector Regression
   Support Vector Regression (SVR) is also one of the most popular algorithms of ML. SVR model is considered as one of the best among ML approaches in terms of accuracy and computational efficiency. SVR auto-tuning mechanism has integrated machine-learning performance model and intelligent search algorithm for an effective exploration of parameter space and efficient training models. The SVR model performance was measured in two programs: sort and wordcount and compared Starfish model. It was shown that the SVR model performance increment was 39% while Starfish model performance increment was 13% when it is analyzed in sort programs. But in wordcount program, the Starfish model performance increment was either similar at 40% or slightly better with a rate of 5% [21].
c. Support Vector Machine
   In another model known as Automated Resource Allocation and Configuration of MapReduce Environment in the Cloud (AROMA) the allocation of resources and configuration of parameters are automated through two-stage optimization framework and ML. This model focused on the way to reducing the cost of big data processing in Hadoop system. The result depicts that the cost of processing 10GB data with AROMA auto-tuning is 36 cents using 5 medium VMs (Virtual Machine) where the one without AROMA auto-tuning is 51 cents using 6 medium VMs. It also shows that resource allocation under AROMA mechanism can cost less compared to the default one. On average AROMA's cost efficiency is 25% [22].

d. K-means++
Unlike the AROMA mechanism, many other mechanisms just focus on improving the performances by reducing the time of data processing. Profiling and Performance Analysis-based System (PPABS) uses two-phase (Analyzer and Recognizer) framework that operates on K-means++ clustering for analyzing and classification approach for recognizing the jobs. The experimental results show that the processing time for Big Data has been reduced in TeraSort and WordCount methods. In an experiment with 10 GB input data set, the usual accomplishment time for TeraSort and WordCount decreases by 38.4% and 18.7% respectively. The reason for this mismatch in performance is characteristics of various jobs [23].

e. Tree-Based Regression
This approach has two phases first one is prediction phase and the second one is optimization phase. In the prediction phase, the performance of MapReduce job is estimated and in the optimization phase, a search for on an average optimum configuration parameter is made by invoking the predictor repeatedly. It is reported that this approach can help the user to increment the performance regarding 2 to 8 times better than prediction phase[24].

2.2.2. Parameters

Above algorithms have been used parameters in Table 1.

Parameters are the main factors that play an important role in Hadoop system for performance improvement. The limitation of default Hadoop system is that the parameters are fixed at default values. Among the 30 effective parameters different models used different parameter configurations. The Table 1. shows some most effective parameters, which were used in optimizing Hadoop performance [25]-[28].

2.2.3. Dataset

Applied above algorithms have been used these datasets. In developing an effective model for improving Hadoop performance, most of the ML algorithm have used multiple datasets in order to make sure that the performance does not decrease against the data size. Random Forest model has used a dataset of 2-5GB for training time and 50GB to 1TB while carrying out the evaluation. The benchmark has been collected from Puma and HiBench suites [1]. SVR based auto-tuning approach has used a dataset of 80GB and 240GB in SNB cluster and applied the sort and wordcount from HiBench benchmark suite [21]. AROMA model has used a dataset of 5GB, 10GB and 20GB and used RandomTextWriter and RandomWriter tools in Hadoop package to produce data of different sizes because of the Sort, Grep and WordCount programs [22]. The PPABS model used 1GB, 5GB and 10 GB datasets. A training set of MR applications consisting of 3 well-known sets namely: Hadoop Examples set, HiBench and Hadoop Benchmarks set (Intel implemented benchmark set) [23].

Table 1. Configuration Parameters in Hadoop system

Parameter	Paper/Reference
io.sort.factor	[1], [21-24]
mapred.job.shuffle.merge.percent	[1], [22], [24]
mapred.output.compress	[1], [24]
mapred.inmem.merge.threshold	[1], [24]
mapred.reduce.tasks	[1], [21], [22]
io.sort.spill.percent	[1], [22-24]
mapred.job.shuffle.input.buffer.percent	[1], [22], [24]
io.sort.record.percent	[1], [22], [24]
io.sort.mb	[1], [22-24]
mapred.compress.map.output	[1]
mapred.tasktracker.map.tasks.maximum	[21], [23]
mapred.tasktracker.reduce.tasks.maximum	[21], [23], [24]
mapred.child.java.opts	[23]
mapred.reduce.parallel.copies	[22], [23]
dfs.block.size	[23]
mapred.map.output.compress	[23], [24]
mapred.job.reduce.input.buffer.percent	[22], [24]
MapHeapSize	[24]
MapTasksMax	[24]

Parameter	Paper/Reference
SplitSize	[24]
HttpThread	[24]
ReduceHeapSize	[24]
ReduceTasksNum	[24]
ReduceCopyNum	[24]
ReduceSlowstart	[24]
JVMReuse	[24]

2.2.4. System configuration

The system configuration has been used in above ML algorithms for improving. The different models have compared the performance with different system. For example, Random- Forest model has compared its outcome against CBO based approach and the configuration has been done in a similar manner also. It has used 10 Sugan servers prepared with Intel-Xeon CPU- E5-2407 2.20GHz and quad-core processor and 32GB PC3 memory connected through gigabit Ethernet [1]. SVR model has used HiBench benchmark for WordCount and Sort benchmark while it has used two clusters namely SandyBridge (SNB) and ZT cluster. SVR performed the experiment on a server which is a dual-core IntelR CoreTM i5-2540M processor running at 2.60GHz and 4GB main memory [21]. On the other, hand SVM model was implemented on settings of 7HP Pro-Liant BL460C G6 blade server together with a HP EVA storage area network that comprised of 10Gbps Ethernet and 8Gbps Fibre/iSCSI dual channels. The small and medium VM (Virtual Machine) used were contained with 1vCPU, 2GB RAM and 50GB hard disk space and 2vCPUs, 4GB RAM and 80GB hard disk space respectively [22]. In addition, K-means++ model has used a cluster that contains five DataNodes and one NameNode. The NameNode and DataNode run on CPU of 2EC2 Compute, and 1EC2 Compute Unit, the memory of 300GB and 200GB respectively [23]. Besides, Tree-Based Regression machine learning algorithm has evaluated its performance on 8 nodes Pdefault setting where each node contains eight Intel i7-4770 cores, 32GB RAM and 2TB disk space [24].

3. PROSPECT OF APPLICATION OF DEEP LEARNING ALGORITHM TO HADOOP PERFORMANCE IMPROVEMENT

Hadoop is an integral part of processing big data. Hadoop performance is an impediment in getting efficient service as the parameters are not self-tuned. Different ML algorithm has been proposed to improve the performances by allowing auto-tuning of the most effective parameters. The analysis of performance with different ML algorithms shows that the self-tuning has improved Hadoop system performance in comparison with default parameter configuration. However, there is a need for a new model to further improve Hadoop system performance with respect to speed and accuracy. Deep learning has been adopted with most popular Theano [29], Tensorflow [30], Cafee [31] library in many sectors of big data processing and it was found to result in improved performance. Deep Learning algorithms are used in processing big data in many giant tech companies including Google, Facebook, Amazon and so on. The authors feel there is a scope for applying deep learning algorithms for self-tuning and improving Hadoop system performance[7].

4. CONCLUSION

In this paper a brief review of the concepts of big data, Hadoop system is presented, self-tuning of Hadoop parameters and ML algorithms. Self-tuning of Hadoop system parameters using ML algorithms has been found to improve performance compared to default parameter configuration. The prospect of the application of deep learning algorithms for self-tuning in Hadoop system to improve speed and accuracy of performance is proposed by the authors.

REFERENCES

[1] Bei Z, Yu Z, Zhang H, Xiong W, Xu C, Eeckhout L and Feng S, 2016, "RFHOC: A Random-Forest Approach to Auto-Tuning Hadoop's Configuration", *IEEE Trans. Parallel Distrib. Syst.*, 27, pp. 1470-1483.
[2] Technology I 2016 Survey on Performance of Hadoop Mapreduce Optimization Methods, 2, pp. 114-21.
[3] Uzunkaya C, Ensari T and Kavurucu Y, 2015, "Hadoop Ecosystem and Its Analysis on Tweets", *Procedia - Soc. Behav. Sci.*, 195, pp. 1890-1897.
[4] Bonifacio A S, Menolli A and Silva F, 2014, "Hadoop MapReduce Configuration Parameters and System Performance: a Systematic Review", *2014 Int. Conf. Parallel Distrib. Process. Tech. Appl.*, pp. 1-7.
[5] Dongyu Feng, Ligu Zhu and Lei Zhang, 2016, "Review of hadoop performance optimization", *2016 2nd IEEE Int.*

Conf. Comput. Commun., pp. 65-68.
[6] Rehan M and Gangodkar D, 2015, "Hadoop, MapReduce and HDFS : A Developers Perspective", *Procedia - Procedia Comput. Sci.*, 48, pp. 45-50.
[7] Oussous A, Benjelloun F, Ait A and Belfkih S, 2017, "Big Data technologies : A survey", *J. King Saud Univ. - Comput. Inf. Sci.*
[8] Jiang D, Ooi B C, Shi L and Wu S, 2010, "The performance of MapReduce", *Proc. VLDB Endow.*, 3, pp. 472-483.
[9] Zhang B, Křikava F, Rouvoy R and Seinturier L, 2015, "Self-configuration of the number of concurrently running MapReduce jobs in a hadoop cluster", *Proc. - IEEE Int. Conf. Auton. Comput,. ICAC 2015*, pp. 149-150.
[10] Lee G J and Fortes J A B, 2016, "Hadoop performance self-tuning using a fuzzy-prediction approach", *Proc. - 2016 IEEE Int. Conf. Auton. Comput. ICAC 2016*, pp. 55-64.
[11] Genkin M, Dehne F, Pospelova M, Chen Y and Navarro P, 2016, "Automatic , on-line tuning of YARN container memory and CPU parameters".
[12] Li C, Zhuang H, Lu K, Sun M, Zhou J, Dai D and Zhou X, 2014, "An adaptive auto-configuration tool for hadoop", *Proc. IEEE Int. Conf. Eng. Complex Comput. Syst. ICECCS*, pp. 69-72.
[13] Kerk C W and Zahid M S M, 2015, "Auto-Tuned Hadoop MapReduce for ECG Analysis", pp. 329-334.
[14] Pospelova M and Affairs P, 2015, "Real Time Autotuning for MapReduce on Hadoop / YARN by Real Time Autotuning for MapReduce on Hadoop / YARN.
[15] Willke T L, 2016, "Gunther : Search-Based Auto-Tuning of MapReduce Gunther : Search-Based Auto-tuning of MapReduce".
[16] Jordan M I and Mitchell T M, 2015, "Machine learning: Trends, perspectives, and prospects", *Science (80-.)*, 349, pp. 255-260.
[17] Qiu J, Wu Q, Ding G, Xu Y and Feng S, 2016, "A survey of machine learning for big data processing", *EURASIP J. Adv. Signal Process.*, 2016, 67.
[18] Tsai C-W, Lai C-F, Chao H-C and Vasilakos A V, 2015, "Big data analytics: a survey", *J. Big Data* 2, 21.
[19] Solanki K and Dhankar A, 2017, "A review on Machine Learning Techniques", 8, pp. 778-782.
[20] Singh Y, Bhatia P K and Sangwan O, 2007, "A review of studies on machine learning Techniques", *Int. J. Comput. Sci. Secur.*, 1, pp. 395-399
[21] Yigitbasi N, Willke T L, Liao G and Epema D, 2013, "Towards machine learning-based auto-tuning of MapReduce", *Proc. - IEEE Comput. Soc. Annu. Int. Symp. Model. Anal. Simul. Comput. Telecommun. Syst. MASCOTS*, pp. 11-20.
[22] Lama P and Zhou X, 2012, "AROMA: Automated Resource Allocation and Configuration of MapReduce Environment in the Cloud", *Proc. 9th Int. Conf. Auton. Comput. - ICAC '12*, 63.
[23] Wu D, 2013, "A Profiling and Performance Analysis Based Self-tuning System for Optimization of Hadoop MapReduce Cluster Configuration".
[24] Chen C O, Zhuo Y Q, Yeh C C, Lin C M and Liao S W, 2015, "Machine Learning-Based Configuration Parameter Tuning on Hadoop System", *Proc. - 2015 IEEE Int. Congr. Big Data, BigData Congr., 2015*, pp. 386-392.
[25] Heger D, 2013, "Hadoop Performance Tuning-A Pragmatic & Iterative Approach", *C. J.*, pp. 1-16.
[26] Devices A M, 2012, "Hadoop Performance Tuning Guide", pp. 1-22.
[27] Shrinivas J, 2011, "Introduction Hadoop Configuration Tuning Best Practices JVM Configuration Tuning OS Configuration Tuning Conclusion & Future Direction", *Technology*, 9, pp. 2011-2011.
[28] Min-Zheng J, 2015, "Research on the performance optimization of hadoop in big data environment", *Int. J. Database Theory Appl.*, 8, pp. 293-304.
[29] The Theano Development Team, "Theano: A Python framework for fast computation of mathematical expressions. Available", arXiv:1605.02688.
[30] Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al., "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems," CoRR, 2016.
[31] Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, Guadarrama S and Darrell T, 2014, "Caffe: Convolutional Architecture for Fast Feature Embedding".

BIOGRAPHIES OF AUTHORS

Md. Armanur Rahman received the B.Sc. degree in computer science and engineering from Asian University of Bangladesh (AUB) in 2010. He is currently working toward the MEngSc degree at the Multimedia University (MMU), Malaysia. His research interest include performance optimization of big data system, data mining, machine learning and image processing.

Dr. Jakir Hossen is graduated in Mechanical Engineering from the Dhaka University of Engineering and Technology (1997), Masters in Communication and Network Engineering from Universiti Putra Malaysia (2003) and PhD in Smart Technology and Robotic Engineering from Universiti Putra Malaysia (2012). He is currently a Senior Lecturer at the Faculty of Engineering and Technology, Multimedia University, Malaysia. His research interests are in the area of Artificial Intelligence (Fuzzy Logic, Neural Network), Inference Systems, Pattern Classification, Mobile Robot Navigation and Intelligent Control.

Dr.Chinthakunta Venkata Seshaiah received his Bachelor of Engineering (B.E.) Degree in Electrical Engineering from S.V. University, Andhra Pradesh, India, in the year 1964.He received Master of Engineering (M.E) degree in High Voltage Engineering from Indian Institute of Science, Bangalore in 1966. He received his Ph.D. degree in Electrical Engineering (in the area of Power Systems) in 1976 from I.I.T. Madras. Later, he worked in the same institute till 2005. He was appointed as Professor of Electrical Engineering in Jan.1993. In 2006, he joined the Faculty of Engineering and Technology, Multimedia University (Melaka) Malaysia and is with them presently as Associate Professor. His research interests are in the areas of Electrical Power Systems, High Voltage Engineering and Instrumentation,Power Electronics and its application to green technology solutions,Electic Power quality improvement and Electrical energy conservation, Power efficient devices and Big data analytics.

Dr. Ho Chin Kuan obtained the B. Sc. (Hons) in Computer Science with Electronics Engineering from University College London, UK. Subsequently, he completed his M.Sc. (IT) and Ph.D. in Information Technology from Multimedia University, Malaysia. At present, he is a Professor and Dean at the Faculty of Computing and Informatics, Multimedia University, Malaysia. His main research interests are Natural Computing, Combinatorial Optimization and Data Mining.

Tan Kim Geok received the B.E., M.E., and PhD. degrees all in elec- trical engineering from University of Technology Malaysia, in 1995, 1997, and 2000, respectively. He has been Senior R&D engineer in EPCOS Singapore in 2000. In 2001–2003, he joined DoCoMo Euro- Labs in Munich, Germany. He is currently academic staff in Multimedia University. His research interests include radio propagation for outdoor and indoor, RFID, multi-user detection technique for multi-carrier tech- nologies, and A-GPS.

Aziza Sultana received the B.Sc. degree in computer science and engineering from Dhaka International University (DIU) in 2016. She is currently searching an oopertunity to continue her higher study. Her research interest include performance optimization of big data system, data mining, machine learning and image processing.

Jesmeen M.Z.H. currently a postgraduate student in Engineering specializing in Artifical Intellegence from Multimedia University (MMU). She completed a bachelor's degree in Computer Science and Engineering from International Islamic University Chittagong, Bangladesh.

Ferdous Hossain received B.Sc. degree in computer science and engineering in 2012 and M.Sc. degree in Information and Communication Technology in 2015 from Mawlana Bhashani Science and Technology University, Bangladesh. Currently, he is pursuing his Ph.D. degree in Faculty of Engineering and Technology at Multimedia University, Malaysia.
His research area includes radio frequency identification (RFID) system, radio propagation for outdoor and indoor, image processing, and big data analysis. He is also an associate member of Bangladesh Computer Society, Bangladesh.