# Big Data Analytics Performance Enhancement For Covid-19 Data Using Machine Learning And Cloud

**[1]Govindaraju G N, [2]Raghavendra B. K., [3]Raghavendra S., [4]Santosh kumar J.**

1Research Scholar, Department of CSE, BGSIT B G Nagar Mandya, VTU belagavi,
govindaraju.gn@gmail.com
[2]Professor and Head Department of CSE, BGSIT B G Nagar Mandya,
raghavendrabk.bgsit@gmail.com
[3]Associate Professor Dept. of CSE, Christ Deemed to be University, Kengeri campus, Bangalore,
raghav.trg@gmail.com
[4]Associate Professor Dept. of CSE, KSSEM Bengaluru, VTU Belagavi
sjankatti@gmail.com

**Article History**: Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 28 April 2021

**Abstract -** The exponential rise in software computing, internet and web-services has broadened the horizon for BigData that demands robust and highly efficient analytics system to serve timely and accurate distributed data support. The distributed frameworks with parallelized computing have been found key driving force behind the contemporary BigData analytics systems; however, the lack of optimal data pre-processing, feature sensitive computation and more importantly feature learning makes major at-hand solutions inferior, especially in terms of time and accuracy. Unlike major at hand methods employing machine learning for BigData analytics, in this paper the key emphasis was made on improving pre-processing, low-dimensional semantic feature extraction and lightweight improved machine learning based feature learning for BigData analytics. Noticeably, the proposed model hypothesizes that an analytics solution with BigData characteristics must have the potential to process humongous, heterogenous, unstructured and multi-dimensional features to yield time-efficient and accuracy analytical outputs. In this reference, we proposed a state-of-art new and robust BigData analytics model, specially designed for Spark distributed framework. To process analytical task our proposed model at first employs tokenization, followed by Word2Vec based semantic feature extraction using CBOW and N-Skip-Gram methods. Our proposed model was found more effective with Skip-Gram Word2Vec feature extraction. Simulation results with a publicly available COVID-19 data exhibited better performance than existing K-Means based MapReduce distributed data frameworks.

**Keywords:** BigData Analytics, Spark Distributed Framework.

## 1. Introduction

The high-pace rise in advanced computing, software technologies, internet and allied information and communication technologies, and web-services has inculcated socio-economic and scientific transition to make timely and optimally accurate decisions [1-3]. The aforesaid technologies have broadened the horizon for the different stakeholders to introduce, explore or exploit, and compute gigantically large data to make realistic decisions pertaining to civic management, business decision, healthcare and decision systems, socialization, scientific and business communication purposes [3-5]. Undeniably, the stakeholder(s) serving different aforesaid purposes often undergo a phase where it requires mining over a gigantically large data, typically called BigData to retrieve the target knowledge or the inference to make decisions [5]. However, mining huge data to retrieve expected information or goal-centric inference is highly complex task [2]. This is because of the heterogenous data nature, with significantly large unstructured and unannotated data condition,etc. [2]. In sync with the motive to develop a robust mechanism to explore, learn and classify a large data, gave rise to a new technology called BigData analytics [2][9]. BigData analytics can be characterized as an advanced computing and knowledge-driven mechanism to explore humongous heterogeneous, unstructured and multi-dimensional data to retrieve target information for real-world decisions [2]. Noticeably, BigData analytics intends to serve the stakeholders timely with optimal accuracy, even under aforesaid heterogeneous, unstructured and multi-dimensional data presence, which is really a challenging task [6-9]. There are many real-time application environments such as Internet-of-Things (IoT), real-time data collection and query driven decision support systems, which demands quick response with optimal accuracy [10]. Such application specific environments demand BigData analytics to have the potential to meet four key demands, often called 4V-conditions; Volume, Variety, Velocity, and Veracity [2]. A typical BigData analytics model requires addressing aforesaid 4V factors where it requires possessing the ability to process a large VOLUME of data with significantly high VARIETIES, within a very small TIME (say, Velocity) without compromising accuracy or VERACITY

[2][11][12][15][16]. Achieving a suitable solution fitting to the aforesaid demands require a robust BigData analytics model, which has alarmed academia-industries to achieve a novel and most-efficient and scalable solution [2][11][12]. A recent study by Gartner has indicated that the BigData would grow 650% in the next five years, and hence alarms the need of a robust analytics to meet distributed decision demand [2][5].

As indicated above, accomplishing a robust BigData analytics model having the potential to process over humongous, heterogenous, unstructured and multi-dimensional data feature require processing elements highly robust in computation, especially in terms of time-efficient data processing. In sync with such needs, different cloud-based analytics models have been proposed, such as Hadoop, Apache Spark, Amazon EC2, etc. [2][13]. However, their efficacy to mine over a gigantically large data with aforesaid features (i.e., unstructured, high heterogeneity, unannotated, and high-dimensional) has always remained a challenge for industries [14][2]. In fact, the majority of the BigData analytics solutions require tailoring large volume of unstructured or semi-structured data into informative feature-instance followed by learning to make target classification [2][13][14]. However, such practices require state-of-art highly robust computing environment. Amongst the major solutions, the parallel computing-based MapReduce and Apache Spark frameworks have turned out well towards BigData analytics [2]. In comparison to the classical parallel processing paradigms such as grid-computing pr Graphical Processing Unit (GPU), MapReduce and Hadoop show two key advantages [2][9][13]. First, its fault-tolerant storage and second high-throughput, where it ensures reliable data processing over multiple clusters and batch-processing. Moreover, the use of Hadoop File Distribution System (HDFS) makes it more scalable, independent and efficient to perform accurate and timely computation [2][13][15][16]. Despite of such efficacy, MapReduce based Hadoop is often criticized for its computational time and cost-overheads for distributed systems [2][31]. Unlike MapReduce-Hadoop system(s), Spark based BigData analytics [2][17][18] have been found more time-efficient due to augmented parallelism and swift computation [17]. Though, Spark has been found more time-efficient and autonomous towards BigData analytics [2], its dependency on data computing environment such as data pre-processing (say, data tailoring), feature learning and classification has remained an open research area [2]. In other words, the efficiency of the aforesaid Spark enabled BigData analytics model is highly dependent on the computing environment and it becomes inevitably vital in case of humongous, heterogenous data with non-linear patterns and large dimensions [2][18]. Additionally, improving time-efficiency often requires BigData analytics model (here, Spark) to have higher level of parallelism [2][19], and therefore inculcating the same has broadened the horizon for academia-industries.

## 2. Related Work

This section discusses some of the key literatures pertaining to machine learning based BigData analytics, parallelization in BigData analytics etc.

BigData being a revolution towards smart thinking, processing and decision making has revitalized human society; however, its resulting efficacy has always been dependent on how better it learns and predicts the targeted outputs [18]. Somewhere it demands analytics to have the potential to process large data and gain optimal knowledge discovery to make optimal decisions [20][19]. Undeniably, BigData has been the core of revolution; however, achieving accurate and timely (processed) knowledge is must [18]. To achieve it, machine learning methods or artificial intelligence seems to be the vital technology [2][20]. In fact, McKinsey believe that the core of BigData analytics is its machine learning driven learning ability and timely decision system [2][20]. In sync with machine learning based BigData both supervised as well as unsupervised learning methods can be viable solution [9][10]. However, under exceedingly dynamic data nature with minimum or even no annotations, unsupervised methods such as K-Means clustering seems to be the better alternative [10]. Though, classification and regression methods such as artificial neural network, support vector machine, logistic/polynomial regression, k-Nearest Neighbour etc. are some of the supervised learning methods used for BigData analytics [2]. On the contrary, unsupervised learning methods such as K-Means are more effective towards contemporary BigData analytics problems, especially for query-driven distributed knowledge services, prediction, and prescription [2]. Unlike major conventional machine learning models, clustering based methods have higher scalability to learn over the large or humongous unannotated data [11]. In the last few years, numerous machine-learning driven BigData analytics models have been proposed such as MapReduce [32], Hadoop [13], however, guaranteeing both computational cost and time-efficiency remained challenge [2][9]. Authors emphasized on applying deep learning concept for BigData analytics; however, failed in addressing key obstacles like unstructured data formats, very swift streaming data, noisy or poor data quality and multi-source data inputs, data imbalance, and unannotated data [16]. To cop-up with realistic analytics demands, authors in [17] suggested to achieve highly scalable architectures, having statistical data analysis capacity even over large or humongous data environment [18]. Though, to improve distributed data platform, authors suggested to use machine learning with vertical and horizontal scaling; however, failed in realising the same with real-world analytics problem. Moreover, the key problem of feature sensitiveness and allied classification problem (data unbalance, curse of dimensionality, unannotated data, etc.) could not be addressed in sync with BigData demands. In [19], authors tried to use Apache Mahout towards humongous online analysis problem. However, could not assess low-dimensional feature sensitiveness which could have yielded more accurate as well as least-exhaustive computing. Authors [20] discussed the different data mining methods for BigData analytics; however, could not classify the methods in terms of corresponding realistic efficacy. Though, it indicates that in addition to the mining methods, clustering too can be a viable solution, especially for large unannotated data. In [11], authors indicated that the use of semantics and application knowledge can help achieving optimal analytics solution, as it can not only reduce computational complexity, but can also reduce time.

## 3. Research Questions

The overall research effort made in this work, intended to achieve a justifiable and suitable answers for the following research questions.

***RQ1:*** *Can the use of data-sensitive pre-processing and low-dimensional feature extraction be effective towards BigData analytics?*

***RQ2:*** *Can the use of word embedding concepts such as Word2Vec be effective to ensure computationally-efficient BigData analytics?*

***RQ3:*** *Can the strategic implementation of heuristic models such as Improved Multi-Objective GA (IMOGA) be efficient towards optimal clustering for BigData analytics?*

***RQ4:*** *Can the realization of IMOGA-K Means clustering algorithm with Spark distributed framework be effective for BigData analytics?*

***RQ5****: What can be the set of computing elements including parallel computing, machine learning and distributed framework to enable optimal BigData analytics?*

Thus, the answer for the above stated questions can eventually contribute a state-of-art new and robust method for BigData analytics serving timely as well as reliable decision support.

## 4. Our Contribution

In majority of the BigData analytics problems, the functional motive used to be either classification, regression or clustering. However, performing feature sensitive clustering with minimum computation time and accuracy has always remained a challenge. On the contrary, clustering based analytics has emerged as inevitable demands to serve business intelligence, research, query-driven distributed data supports etc. Most of the at-hand analytics models demand timely and reliable analytics output, which is undeniably very challenging especially under humongous, heterogenous, unstructured data nature. Thus, realizing the inevitable significance of a lightweight, time-efficient and accurate BigData analytics model, in this research paper the emphasis is made on enhancing each comprising step including data pre-processing, feature extraction, feature sensitive clustering and cluster optimization. In our proposed BigData analytics model we focused on exploiting the efficacy of the different technologies such as machine learning, semantic feature embedding, evolutionary computing and Apache Spark distributed framework to design a state-of-art new and robust BigData analytics model. Noticeably, being a BigData analytics problem, which is expected to undergo 4V-specific demands, we designed our proposed analytics model in such manner that it could address all allied aspects including large volume of data, multi-dimensional features, unstructured data etc., while accomplishing timely (i.e., signifying velocity), and accurate (say, veracity) analytics outcome. The strategic implementation of the overall proposed model encompassed the following key phase-wise processing elements.

1. *Data Collection*
2. *Pre-processing or Tokenization*
3. *Latent Semantic Feature Extraction*

## 5. Results and Discussions

Realizing the fact that the majority of the contemporary BigData analytics model either applies classification systems, regression or the clustering models to perform data analysis and resulting prediction or prescription tasks. Majority of the distributed frameworks, especially designed towards BigData analytics tasks employ advanced computing environment such as machine learning or artificial intelligence paradigms. However, the high-pace up surge in data heterogeneity, unstructured data and allied high-dimensional complexity confine classical distributed frameworks. To cope up with the increasing humongous data and allied complexity demands more effective computing environment to process raw gigantic inputs (i.e., VOLUME) to mine over the large features (i.e., VERIETY) so as to achieve inferable or significant analytics output (i.e., VERACITY), even in small time-period (i.e., VELOCITY). In sync with these goals, BigData frameworks such as Hadoop, Apache Spark, Amazon EC2 etc. have been proposed. However, ensuring eventual analytics goal under contemporary data condition has always remained the challenge. Though, a large study has been done towards MapReduce (Hadoop distributed framework); however, it has been found more time-exhaustive than the Apache

Spark. Considering it as motivation, in this research paper the key emphasis was made on designing a state-of-art new and robust heuristically transitioned and semantic feature driven clustering environment was designed for BigData analytics. Unlike the major existing researches where authors merely focused on manipulating the spark framework with different node/cluster configuration to improve the performance, we focused on making a complete intrinsic functional as well as architectural transition. To achieve it, we made value addition and enhancement at the different layers of the analytics design, including data collection and pre-processing, latent semantic feature extraction followed by a state-of-art new heuristically driven K-Means clustering development. In sync with typical BigData problem, where the data could be in humongous size, encompassing exceedingly high heterogeneity and multi-dimensional features, at first, we focused on data pre-processing over the input raw data. Noticeably, in this research we considered COVID-19 datasets [82], as case study where the key motive was to cluster the input data into different types of contents based on similarity, correlation and similar latent-information. The considered task represents a typical text classification problem, where the intended BigData analytics model is expected to cluster the large inputs based on corresponding feature-similarity. However, unlike classical clustering problem, the at hand problem (i.e., COVID-19 dataset or CORD-19) is highly complex due to large unannotated and unstructured data. Noticeably, the considered case study (i.e., CORD-19 or COVID-19 database analysis) specially demands a robust and highly accuracy BigData analytics solution which could learn over a gigantically large data (Note: the proposed CORD-19 data encompasses almost 4.5 Lakhs of scholarly articles pertaining to COVID-19 pandemic) and cluster them as per the feature similarity. The considered dataset encompasses scholarly or scientific articles containing Coronavirus, probable reasons, symptoms, demographics and infection related probability, current trends, future predictions, vaccines, testing methods and their efficacy, false positive cases etc. It indicates the highly heterogenous and multidimensional data nature. On the contrary, a total of 4.5 Lakhs of articles tells the story of a gigantic dataset. The key problem with this dataset was unstructured humongous size and more importantly extracting the features even with low dimension, as merely applying word-matching concept would impose huge computational overheads and hence reduced performance. Moreover, clustering the articles based on the diverse features (due to multiple similar or highly close features) was a real-challenge. To alleviate it, we designed a state-of-art new heuristically driven clustering model.

| NO. OF CLUSTERS | SILHOUETTE COEFFICIENT CBOW |
|---|---|
| **2** | **0.821** |
| **4** | **0.887** |
| **6** | **0.953** |
| **8** | **0.921** |
| **10** | **0.948** |
| **12** | **0.961** |
| **14** | **0.970** |

The above result (Table I) shows the clustering efficacy assessment in terms of Silhouette coefficient, where higher value of the coefficient signifies that the clustering has been done better or efficiently (signifying, VERACITY). As already discussed in the precious sections, the proposed model extracted Word2Vec features for both CBOW as well as SG (with N=1). Therefore, we examined the performance in terms of the Silhouette coefficient over the different clusters. Noticeably, here, cluster states the initial chromosome size defined by user stating that it could be the probable size of the cluster, which is yet to be processed and optimized by our proposed IMOGA heuristic model. In other words, in our proposed IMOGA K-Means clustering algorithm, a user can define a tentative number of clusters (say, initial cluster), with reference to which IMOGA starts optimization and based on associated (data element's) feature it maps each data element to the most suitable cluster and thus optimizes the cluster. To assess the performance, we varied expected number of clusters and examined the Silhouette coefficient obtained over the different inputs. As indicated in Table I, with increase in cluster size, the Silhouette coefficient too increases. Taking an average of the Silhouette coefficients, we found that the mean coefficients are 0.946 for SG (n=1) and 0.923 for CBOW., which is in close vicinity to the setup with cluster-size 6. It indicates that under proposed data condition, the optimal number of clusters seems to be 6 that can assure providing optimal clustering performance with SG features. Moreover, the results affirm (Table I) that in comparison to SG features, the use of SG seems more significant because of its better feature (even with reduced number of feature vectors) presentation (and hence has higher Silhouette coefficient with IMOGA K-Means clustering).

| NO. OF NODES | CBOW<br>EXECUTION TIME (SECONDS) |
|---|---|

| | |
|---|---|
| **1** | **31.4** |
| **2** | **27.6** |
| **3** | **25.3** |
| **4** | **24.7** |
| **5** | **22.8** |
| **6** | **22.1** |
| **AVERAGE** | **25.65** |

Observing the results (Table IV), it can easily be found that the proposed IMOGA K-Means model with SG features exhibits lower average computation time (25.65 second) than CBOW features (25.35 seconds). Interestingly, the disparities amongst the execution times with the different number of nodes are very small, signifying that the proposed IMOGA K-Means based BigData analytics takes almost similar execution with the different nodes; however, the accuracy with large cluster is more with SG features. Therefore, the proposed model infers that the proposed IMOGA K-Means clustering with SG features can be more suitable to serve as BigData analytics.

Recalling the previous discussion that most of the heuristic methods including GA might undergo local minima and convergence problem, which can be more severe with humongous, heterogenous, unlabeled, high-dimensional data, we examined fitness achievement with the different feature sets. To achieve it, we applied IMOGA K-Means clustering algorithm with both CBOW and SG features. To test convergence probability, we considered smaller number of generation (Fig. 1 and Fig. 2). Here, we considered the total number of generations as 10 and estimated fitness value over CBOW (Fig. 1) and SG features (Fig. 2). Observing the results (Fig. 1 and Fig. 2), it can easily be found that the proposed SG feature (with n=1) achieves higher fitness value swift and therefore can avoid a large redundant computation even over large search space. On the other hand, the fitness estimation over CBOW (Fig. 1) indicates that due to relatively large semantic features learning over it turns out to be difficult, which could be alleviated by taking smaller window size. Note, we had considered the corpus size as 500, while the window size was considered as 5. On the contrary, SG method applied N=1, and therefore learning over more segmented features in SG is easier. Consequently, SG feature with N=1 exhibited better; however, it might be computationally heavier, which can be reduced by taking N=2, 3, or 6 (or higher value).

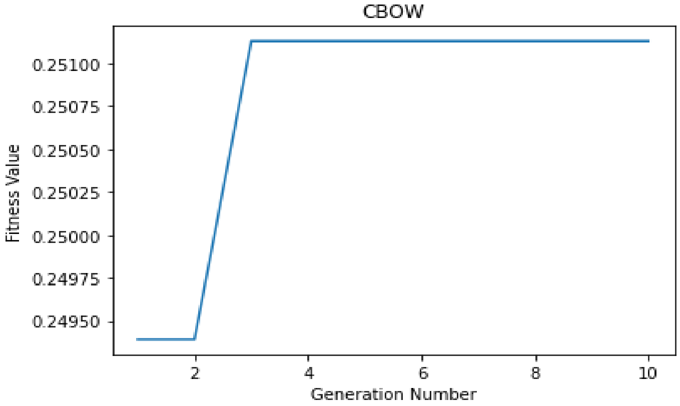

Fig. 1 Fitness value Vs. Generations in CBOW feature learning and clustering

## 6. Conclusion

This research work primarily focused on developing an enhanced distributed framework for BigData analytics, especially under typical large data with heterogenous, unstructured a multi-dimensional feature. Realizing the fact that merely employing distributed framework such as MapReduce and Spark can't yield optimal performance, until it doesn't address the key issues of data-heterogeneity, humongous size with unannotated data structure etc. In sync with this inference, this research proposed a state-of-art new and robust Spark distributed framework was developed with semantic word-embedding, and evolutionary computing assisted lightweight clustering algorithm. Here, the key intend was to enhance data quality and analytics-oriented suitability followed by time-efficient and accurate clustering to make optimal decisions. To achieve it, the proposed model at first tokenized the input data, which is a COVID-19 related data repository containing almost 4.5 lakhs of articles and statistical document related to COVID-19's symptoms, vaccines, current measures, health complications etc. Once tokenizing the inputs, a relevant dictionary was constructed. Subsequently, for each data instance, word-embedding methods CBOW and N-Skip Gram were applied to extract the feature. Noticeably. The parallelized implementation of SG driven IMOGA K-Means enabled time-efficient analytics to serve real-world demands. Interestingly, with higher clusters and node configuration the proposed distributed framework exhibited better time efficiency. Moreover, higher Silhouette coefficient affirms optimality of clustering and hence its reliability towards real-world analytics purposes. Noticeably, the proposed Spark BigData analytics model was applied to perform content prediction and prescription over COVID-19 pandemic related data, and therefore it can be vital for the different tasks related to vaccine related research, query-driven data retrieval and visualization etc. However, the proposed model can be applied for any BigData analytics problems including text analytics, heterogenous data processing and allied analytics tasks.

**REFERENCES**

1. R. Krikorian. (2010). Twitter by the Numbers, Twitter. [Online]. Available: http://www.slideshare.net/raf_krikorian/twitter-by-the-numbers?ref=http://techcrunch.com/2010/09/17/twitter-seeing-6-billion-api-callsper-day-70k-per-second/
2. A. L’ Heureux, K. Grolinger, H. F. Elyamany and M. A. M. Capretz, "Machine Learning with Big Data: Challenges and Approaches," in IEEE Access, vol. 5, pp. 7776-7797, 2017.
3. ABI. (2013). Billion Devices Will Wirelessly Connect to the Internet of Everything in 2020, ABI Research. [Online]. Available: https://www.abiresearch.com/press/more-than-30-billion-devices-willwirelessly-conne/
4. W. Raghupathi and V. Raghupathi, ``Big data analytics in healthcare: Promise and potential," Health Inf. Sci. Syst., vol. 2, no. 1, pp. 1_10, 2014. [4] O.Y. Al-Jarrah, P. D. Yoo, S. Muhaidat, G. K. Karagiannidis, and K. Taha, ``Efficient machine learning for big data: A review," Big Data Res., vol. 2, no. 3, pp. 87_93, Sep. 2015.
5. M. A. Beyer and D. Laney, “The importance of `big data': A definition," Gartner Research, Stamford, CT, USA, Tech. Rep. G00235055, 2012.
6. V. Mayer-Schönberger and K. Cukier, Big Data: A Revolution That Will Transform How We Live, Work, and Think. Boston, MA: Houghton Mifflin Harcourt, 2013.
7. H. V. Jagadish et al., ``Big data and its technical challenges," Commun. ACM, vol. 57, no. 7, pp. 86_94, 2014.
8. M. James, C. Michael, B. Brad, and B. Jacques, Big Data: The Next Frontier for Innovation, Competition, and Productivity. New York, NY: McKinsey Global Institute, 2011.
9. J. Singh, "Real time BIG data analytic: Security concern and challenges with Machine Learning algorithm," 2014 Conference on IT in Business, Industry and Government (CSIBIG), Indore, India, 2014, pp. 1-4.
10. M. Mohammadi, A. Al-Fuqaha, S. Sorour and M. Guizani, "Deep Learning for IoT Big Data and Streaming Analytics: A Survey," in IEEE Communications Surveys & Tutorials, 2018, vol. 20, no. 4, pp. 2923-2960.
11. P. Bellini, F. Bugli, P. Nesi, G. Pantaleo, M. Paolucci and I. Zaza, "Data Flow Management and Visual Analytic for Big Data Smart City/IOT," 2019 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation, Leicester, United Kingdom, 2019, pp. 1529-1536.
12. T. Chardonnens, “Big Data analytics on high velocity streams: specific use cases with Storm”, Software Engineering Group, Department of Informatics, University of Fribourg, Switzerland, 2013.
13. E. A. Mohammed, B. H Far, and C. Naugler, “Applications of the MapReduce programming framework to clinical big data analysis: current landscape and future trends”, BioData Mining 2014, pp. 7:22.
14. I. A. Ajah, H F. Nweke, “Big Data and Business Analytics: Trends, Platforms, Success Factors and Applications”, Big Data and Cognitive Computing, 2019, 3, pp. 1-32.
15. R. Narasimhan and T. Bhuvaneshwari, ``Big data: A brief study," Int. J. Sci. Eng. Res., vol. 5, no. 9, pp. 350_353, 2014.
16. F. J. Ohlhorst, Big Data Analytics: Turning Big Data into Big Money, vol. 15. Hoboken, NJ: Wiley, 2012.
17. A. Kaplunovich and Y. Yesha, "Consolidating billions of Taxi rides with AWS EMR and Spark in the Cloud: Tuning, Analytics and Best Practices," 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 2018, pp. 4501-4507.
18. M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica, ``Spark: Cluster computing with working sets," in Proc. 2nd USENIX Conf. Hot Topics Cloud Comput., 2010, p. 10.
    C.-T. Chu et al., ``Map-reduce for machine learning on multicore," in Proc. 20th Conf. Adv. Neural Inf. Process. Syst. (NIPS), 2006, pp. 281-288.

19. A. C. Onal, O. BeratSezer, M. Ozbayoglu and E. Dogdu, "Weather data analysis and sensor fault detection using an extended IoT framework with semantics, big data, and machine learning," 2017 IEEE International Conference on Big Data (Big Data), Boston, MA, USA, 2017, pp. 2037-2046.