International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-8 Issue-11, September 2019

The Deep Learning and Apache Spark Enabled Architecture for Improving the Performance of Big Data Classification

Anilkumar V. Brahmane, B. Chaitanya Krishna

Abstract*: *At present the Big Data applications, for example, informal communication, therapeutic human services, horticulture, banking, financial exchange, instruction, Facebook and so forth are producing the information with extremely rapid. Volume and Velocity of the Big information assumes a significant job in the presentation of Big information applications. Execution of the Big information application can be influenced by different parameters. Expediently search, proficiency and precision are the a portion of the overwhelming parameters which influence the general execution of any Big information applications. Due the immediate and aberrant inclusion of the qualities of 7Vs of Big information, each Big Data administrations anticipate the elite. Elite is the greatest test in the present evolving situation. In this paper we propose the Big Data characterization way to deal with speedup the Big Data applications. This paper is the review paper, we allude different Big information advancements and the related work in the field of Big Data Classification. In the wake of learning and understanding the writing we discover the holes in existing work and techniques. Finally we propose the novel methodology of Big Data characterization. Our methodology relies on the Deep Learning and Apache Spark engineering. In the proposed work two stages are appeared; first stage is include choice and second stage is Big Data Classification. Apache Spark is the most reasonable and predominant innovation to execute this proposed work. Apache Spark is having two hubs; introductory hubs and last hubs. The element choice will be occur in introductory hubs and Big Data Classification will happen in definite hubs of Apache Spark.

Keywords- Big Data Classification, Deep Learning, Apache Spark

I. INTRODUCTION

The term information mining alludes to the non-trifling extraction of legitimate, certain, conceivably valuable and at last reasonable data in huge databases with assistance of the rising registering advancements [7] [6]. As informational indexes become bigger and progressively confused, an outrageous learning machine (ELM) that keeps running in a customary sequential condition can't understand its capacity to be quick and successful [2]. Viable Big Data Mining requires adaptable and proficient arrangements that are additionally effectively available to clients at all degrees of mastery [20].

Manuscript published on 30 September 2019.
***Correspondence Author(s)**
Anilkumar V. Brahmane*, Research Scholar, Department of Computer Science and Engineering, KL Deemed to be University, Vijaywada, A.P., India. Email: vb_anil@yahoo.co.in
Dr. B. Chaitanya Krishna, Professor, Department of Computer Science and Engineering, KL Deemed to be University, Vijaywada, A.P., India.

Huge Data is "high-volume, high-speed, or potentially high assortment data resources that require new types of preparing to empower upgraded basic leadership, knowledge revelation and procedure enhancement" [1]. By and by, the term alludes to datasets that are progressively hard to gather, precise and procedure utilizing conventional approachs. Huge Data is clear, empowering associations to pick up bits of knowledge from new wellsprings of information that were ever mined in past. Enormous information innovation was generally connected. In the scholastic network, the regarded diaries "Nature" and "Science" have separately propelled huge information issues named "Enormous Data" and "Manage Data", which examine an assortment of issues experienced in huge information innovation from the Internet innovation, financial matters, supercomputing, organic sciences, drug and numerous different viewpoints [4]. The monstrous volume of data assembled by contemporary frameworks wound up inescapable, the same number of research exercises require gathering progressively immense measures of information [1]. All are functioning admirably at little information however their exhibition decays with the expansion in the size of the information. Handling information, which changes in volume, assortment and veracity, is past the limit of run of the mill database programming to imprison, heap, control and assess [6]. A parallel figuring stage has been broadly connected in enormous content information and chart information mining [22]. The ongoing prominence of these parallel programming models has summoned huge enthusiasm for actualizing adaptable renditions of Machine Learning calculations. Normally called enormous information handling, examining huge volumes of information from topographically circulated locales with AI calculations has risen as a significant investigative device for governments and global companies. AI methods figure out how to comprehend complex informational indexes to settle on basic choices and tune the highlights for abstracting superior. The exhibition expected depends personally on the model highlights fundamental the parallel programming structure [24]. The conventional knowledge requires the gathering of the considerable number of information over the world to a focal server farm area, to be handled utilizing information parallel applications. [5]. Hence, we need the unique instruments for enormous information. For our situation, we are investigating the apache flash as the huge information examination apparatus for security examination [21]. MapReduce and Spark are huge scale stages supporting a few calculations, for example, hereditary calculations and molecule swarm improvement, among others [3] [13].

MapReduce is a proficient parallel programming model for the information escalated applications for handling enormous informational indexes. Inferable from couple of inadequacies in MapReduce new model are exceptionally versatile. MapReduce utilizes a Hadoop circulated document framework for parallel preparing. MapReduce model, for example, high plate peruses and composes, low through put brings about low execution while dealing with groups. In the interim the express advancement in innovation rose stages like Kafka, Spark, Flume and some more. In light of the disadvantages in MapReduce up and coming advancements supplanted it [24]. Apache Spark is a quick bunch processing motor that gives comparable adaptability and adaptation to non-critical failure properties to MapReduce, however rather than Hadoop's two-organize plate based Map Reduce worldview [17].

Apache (Spark) is an open source parallel information preparing system, pulling in high consideration in the field of enormous information investigation and man-made consciousness. While Spark essentially holds information in memory during calculation, plates on laborer hubs are utilized for the mix activity and client's predefined storing activity [11]. The main role of Spark is to effectively deal with information of an enormous size. Utilizing worked in libraries independently, existing frameworks have figured out how to utilize Spark for such examination of Big Data [16]. There are as of now two LDA frameworks on Spark, to be specific, Spark LDA [19] and the official one in ML lib [20] (which uses variational induction as opposed to CGS). In any case, these frameworks perform and scale ineffectively, since them two have high computational expense [9]. Sparkle is a quick and flexible bunch registering stage planned huge scale information handling. Flash can perform bunch handling in a base unit, called smaller scale cluster process. Since calculation is performed on-memory in the Spark stage, the presentation of information handling is improved maintaining a strategic distance from costly circle get to. The Spark undertaking comprises of different segments firmly coupled together, for example, Spark center, Spark SQL, MLlib. The Spark center incorporates Spark's essential capacities, for example, task booking, memory the executives, fiasco recuperation, and cooperation with the capacity framework [19].

The fundamental commitments of this paper are as per the following.

1. The writing study is done to discover the holes in the current work and approachs which are proposed and executed by different writer's
2. The new and novel methodology for Big Data Classification is proposed.

The structure of this paper is as per the following.

Initially, in the area II writing study is completed, identified with Big Data Classification . At that point in area III the difficulties in existing framework and holes are discover. In area IV the targets are notice and new novel methodology for Big Data arrangement is proposed. At last in area V finish of the work is given.

II. LITERATURE SURVEY AND RELATED WORK.

In this area we have learn and comprehend the different existing approachs proposed by many author's. Subsection An is concentrating on Big Data Classification. Subsection B is concentrating on Parallel preparing and it's structure. Though subsection C concentrating on Machine learning approach on information mining. This investigation gives us the enough learning and data about the current situation identified with Big Data , Data Mining and Classification and parallel structures.

A. *Enormous Data Classification*

In this area different existing approachs for Big Data Classification is talked about.

Closest Neighbor Classification for High-Speed Big Data Streams Using Spark [1] by the S. Ramírez-Gallego proposed a philosophy a proficient gradual example choice strategy for enormous information streams. The creators proposed DS-RNGE – Distributed Relative Neighborhood Graph Edition incorporates an occurrence choice method that always improves the exhibition. The impediments of this technique is enormous size and excess issues were not taken care of by the strategy.

A Parallel Multiclassification Algorithm for Big Data Using an Extreme Learning Machine [2] by M. Duan proposed a philosophy a proficient ELM dependent on the Spark system (SELM). The creator proposed a SELM – Spark Extreme Machine Learning which parts the informational index sensibly, which makes numerous calculations to be performed locally as could be allowed and brings down the correspondence cost and I/O cost. Furthermore, SELM has the most elevated speedup. The confinements of this technique is more memory and preparing units are required, to empower the parallel calculations to increase a superior speedup.

Enormous Scale Machine Learning dependent on Functional Networks for Biomedical Big Data with High Performance Computing Platforms [3] by Elsebakhi, E proposed a system utilitarian systems dependent on penchant score and Newton Raphson-most extreme probability enhancements as another huge scale AI classifier. The athor proposed technique which spares the computational time and have dependable exhibitions alongside future road for augmentation to manage cutting edge sequencing information on superior registering stages. The confinements of this technique is Performance was better with insignificant estimations of right arrangement rate (CCR). An Ensemble Random Forest Algorithm for Insurance Big Data Analysis [4] by W. Lin proposed an A group arbitrary woods calculation which utilized the parallel registering capacity and memory-store system improved by Spark. The proposed technique The group arbitrary backwoods calculation outflanked SVM and other characterization calculations in both execution and exactness utilizing the imbalanced information, and it is helpful for improving the precision of item showcasing contrast with the customary fake methodology. The restrictions of this technique is it required to improve the exactness of expectation dependent on enormous information examination for which the future expansion might be founded on any profound learning calculations.

Utilizing AI to enhance parallelism in enormous information applications [5] by Hernández proposed an approach utilizing AI to enhance parallelism in huge information. The proposed strategy can precisely anticipate the execution time of a major information based application with various record sizes and parallelism settings utilizing the correct models. The restrictions of strategy is endured because of multifaceted nature and required the brought together administration of situations with different advancements.

A Distributed Evolutionary Multivariate Discretizer for Big Data handling on Apache Spark [6] by Ramírez-Gallego. The creator proposed a system A dispersed discretization calculation for Big Data investigation dependent on developmental enhancement. The proposed approach is having beneficial on account of the effortlessness and exactness pace of the created arrangements utilizing a few parameters, this was great. The confinement of the technique is in gushing condition where discretization plans develop after some time, idea floats may influence them.

Mining Maximal Frequent Patterns in Transactional Databases and Dynamic Data Streams: a Spark-based Approach [7] is proposed by Karim. The creator proposed a strategy - The ASP-Tree Construction Algorithm, database Transformation, unraveling Techniques and preparing pipeline. The technique is having the favorable circumstances like it decrease both hunt space, mining and looking through time. The confinement of the strategy incorporate the accompanying: Firstly, it's hard to produce tremendous engineered dataset. Furthermore, genuine value-based datasets are once in a while discovered uninhibitedly open or with information permit understanding.

ZenLDA: Large-Scale Topic Model Training on Distributed Data-Parallel Platform [8] is proposed by B. Zhao. The creator proposed a procedure LDA-Latent Dirichlet Allocation preparing framework named ZenLDA. ZenLDA, pursues a summed up plan for the disseminated information parallel stage. The curiosity of ZenLDA(1) it changes over the normally utilized sequential Collapsed Gibbs Sampling (CGS) induction calculation to a Monte-Carlo Collapsed Bayesian (MCCB) estimation technique, which is embarrassingly parallel; (2) it breaks down the LDA derivation equation into parts that can be tested all the more effectively to lessen calculation multifaceted nature; (3) it proposes a conveyed LDA preparing system, which speaks to the corpus as a coordinated diagram with the parameters explained as comparing vertices and actualizes ZenLDA and other surely understood deduction strategies dependent on Spark. The fundamental points of interest in strategy that it demonstrated great versatility when managing enormous scale theme models on the information parallel stage. The conveyed information parallel reflection (particularly diagram deliberation) isn't just achievable and helpful yet in addition effective and adaptable. The restrictions of the technique is the auto-tuning of the framework parameters is the real challenge as it depends on the experience and testing.

B. *Parallel processing framework- Apache Spark , Map-Reduce .*

When it comes the principal appropriated systems for colossal scale information preparing the main name come infornt and that is Google, which is in charge of structured the Map Reduce in 2003 [29]. Guide Reduce utilizes the groups of PCs are utilized for consequently handling information. Mappers and Reducers are the two parameters that client need to actualize in Map Reduce. In Map Phase the key – esteem sets are perused legitimately from disseminated record framework. These are change into another arrangement of sets. Hub are perusing and changing a lot of sets from at least one information segments. In Reduce Phase client characterized capacities are utilized to sent the key incidental combines and converged to get the last yield. For more data adjoin Map Reduce and others circulated structures, if it's not too much trouble check [30].

Another prominent open source usage of Map Reduce is Apache Hadoop [31] [32]. Its dependable, versatile and appropriated figuring. Hadoop is having a restrictions that is isn't wel appropriate for where there is requirement for unequivocal information reusage. For instance online intelligent, or potentially iterative figuring are influenced by this issue [33].

Apache Spark is a quick and universally useful group registering framework. It gives abnormal state APIs in Java, Scala, Python and R, and an advanced motor that supports general execution diagrams. It likewise bolsters a rich arrangement of higher-level apparatuses including Spark SQL for SQL and organized information processing, MLlib for machine learning, GraphX for chart preparing, and Spark Streaming.

Versatile Distributed Datasets (RDD) is a principal information structure of Spark. It is a permanent disseminated gathering of articles. Each dataset in RDD is isolated into coherent segments, which might be figured on various hubs of the bunch. RDDs can contain any kind of Python, Java, or Scala objects, including client characterized classes. RDD is a perused just, parceled gathering of records. RDDs can be made through deterministic tasks on either information on stable stockpiling or different RDDs. RDD is an issue tolerant gathering of components that can be worked on in parallel.

There are two different ways to make RDDs − parallelizing an existing accumulation in your driver program, or referencing a dataset in an outside capacity framework, for example, a common record framework, HDFS, HBase, or any information source offering a Hadoop Input Format.

Sparkle utilizes the idea of RDD to accomplish quicker and proficient MapReduce activities.

The key thought of sparkle is Resilient Distributed Datasets (RDD); it underpins in-memory preparing calculation. This implies, it stores the condition of memory as an article over the occupations and the item is sharable between those employments. Information partaking in memory is 10 to multiple times quicker than system and Disk.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

Flash likewise enables us to the RDD 's API in gushing condition through the change of information streams into little clusters. Flash Streaming's plan empowers a similar group code to be utilized in gushing investigation, without a necessity for huge adjustments.

The AI library of Spark is having sevral bundles in MLlib which incorporate learning calculations and utilities [34] [35]. Characterization, advancement, relapse, cooperative separating, bunching and information pre-preparing are the different assignments can be done on this Mllib.

C. *Machine Learning Approaches for Data Mining*

AI calculation are chipping away at tremendous volume of information. Speed of the information age additionally assumes the significant job where AI calculation are essential. With the headway in tactile, computerized capacity, and web and correspondence advances, ordinary ML innovative work - which exceed expectations in model, calculation and hypothesis developments are presently tested by the developing commonness of enormous information accumulations, for example, many hours video-sharing locales consistently, or petabytes of web based life on billion or more client interpersonal organizations. The ascent of huge information is additionally being joined by expanding craving for higher dimensional and progressively complex ML models with billions to trillions of parameters. So as to help the consistently expanding multifaceted nature of information, or to acquire still higher prescient accuracy(e.g. for better client administration and restorative conclusion) and bolster increasingly wise tasks(e.g. driver less autos and semantic translation of video information). Preparing such enormous ML models over such huge information is past the capacity and calculation abilities of single machine. This hole has roused a developing collection of late work in disseminated ML, where ML projects are executed crosswise over research groups, server farms, and cloud suppliers mind tens to thousands of machine.

D.*Profound Learning*

Profound taking in systems are recognized from the more ordinary single-covered up layer neural systems by their profundity; that is, the quantity of hub layers through which information must go in a multistep procedure of example acknowledgment.

Prior variants of neural systems, for example, the first perceptrons were shallow, made out of one info and one yield layer, and at most one concealed layer in the middle. Multiple layers (counting info and yield) qualifies as "profound" learning. So profound isn't only a popular expression to cause calculations to appear as though they read Sartre and tune in to groups you haven't knew about yet. It is a carefully characterized term that implies more than one concealed layer.

In profound learning systems, each layer of hubs prepares on a particular arrangement of highlights dependent on the past layer's yield. The further you advance into the neural net, the more perplexing the highlights your hubs can perceive, since they total and recombine highlights from the past layer.

III. CHALLENGES IN EXISTING METHODOLOGIES AND RESEARCH GAP

A.*Challenges in existing methodologies.*

• Apache Spark, famously known for enormous information handling capacity, is a dispersed open-source stage that uses the idea of disseminated memory to encourage huge information preparing capably. From the part of execution, it is as yet a major test to acquire the best yield from Spark, since the Spark setup settings with enormous parameters design influence its exhibition everywhere degree [23].

• The fast advancement of innovation has prompted age of enormous scale information, and it has arrived at a point where consecutive and conventional information preparing model are not ready to give all the expository and computational arrangements. Henceforth, to conquer these difficulties various bunches and conveyed registering methodologies and structures were advanced to help enormous scale information escalated applications [25].

• Programmable Clusters condition has brought a few difficulties: Firstly, numerous applications should be modified in a parallel way, and the programmable Clusters need to process more sorts of information figuring; Secondly, the adaptation to internal failure of the Clusters is progressively significant and troublesome; Thirdly, Clusters powerfully arrange the registering assets between shared clients, which builds the obstruction of the applications. With the quick increment of utilizations, Clusters figuring requires a working answer for suit various computations [18].

• Common difficulties during information change and highlight extraction include: Taking absolute information, (for example, nation for geolocation or classification for a motion picture) and encoding it in a numerical portrayal. Separating helpful highlights from content information, Dealing with picture or sound information [15].

• Achieving an information model completely perfect for Spark and MPI that gives versatility, execution and interoperability appropriate for logical information absorption stays a test not completely fulfilled by any current stage, and this is the objective of our structure [12].

• The rising mechanical enormous information have '5V' qualities (volume, speed, assortment, veracity, and worth), which difficulties the customary prognostics models [10].

B.*Research Gaps*

In the wake of directing a thorough writing survey, we discovered research hole. For Big information order different strategies and systems are utilized yet there are a few impediments in every strategy.

This constraints causes us to discover the exploration holes in existing frameworks.

In section II we conduct the literature review, following are the gaps that we noticed in the given methodologies proposed in literature.

1. Large size and redundancy problems were not handled by the method.
2. More memory and processing units are required, to enable the parallel algorithms to gain a better speedup.
3. Performance was better with minimal values of correct classification rate (CCR).

4. Required to improve the accuracy of prediction based on big data analysis for which the future extension may be based on any deep learning algorithms.
5. Suffered due to complexity and required the centralized management of environments with multiple technologies.
6. In streaming environment where discretization schemes evolve over time, concept drifts might affect them.
7. It's difficult to generate gigantic synthetic dataset.
8. Real transactional datasets are rarely found freely open or with data license agreement.
9. The auto-tuning of the system parameters is the major challenge as it is based on the experience and testing.

IV. RESEACH OBJECTIVES AND PROPOSED METHODOLOGY

A. *Research Objectives*

The goal of research undertaking condenses what is to be accomplished by the examination. The examination goals are the particular achievements the specialist would like to accomplish by the investigation. A plainly characterized research target will assist the specialist with focusing on the examination.

Taking into consideration the above definition about research objectives, we frame out the following research objectives. These objectives are depend upon the challenges and research gap which we identify in section III.

1. To develop an effective training algorithm for the deep learning networks to improve the classification performance.
2. To devise a new deep learning classifier tending to contribute highly accurate classification.
3. To model a feature vector comprising of initial nodes and final nodes on the basis of spark architecture for classification.
4. To devise an algorithm that is capable for tuning the deep learning network to generate optimal weights.
5. To design a hybrid optimization algorithm for selecting the optimal features for effective classification.

B. *Proposed methodology*

The essential expectation of this paper is to structure and build up a procedure for enormous information order utilizing the sparkle engineering. The proposed strategy for enormous information arrangement will incorporate two procedures: highlight determination and huge information order. By and large, the flash engineering is created with the underlying hubs and last hubs where the component choice and arrangement stage advances.

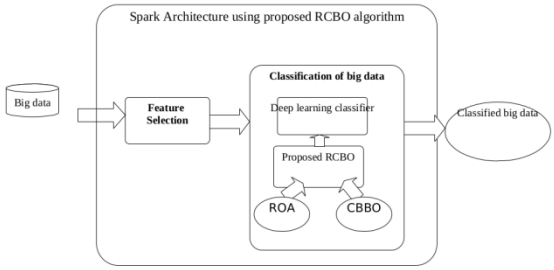
At first, the enormous information characterization is started utilizing the component determination venture in the underlying hubs of the Spark system. The underlying hubs of Spark will peruse the huge information from different dispersed frameworks and will shape the component vector dependent on the created streamlining calculation that will choose the ideal highlights for order. The chose highlights will be encouraged to the last hubs of sparkle that will be furnished with the profound learning systems. The ideal huge information order will be performed at the last bunch hubs of sparkle with the end goal that the proposed enhancement based profound learning systems will be in charge of the ideal information characterization.

In the last hubs of Spark, the profound learning systems will be tuned ideally utilizing the proposed Rider Chaotic Biography improvement (RCBO) calculation. The proposed Rider Chaotic Biography improvement (RCBO) calculation will be the reconciliation of the Rider Optimization Algorithm (ROA) [27] and the standard disorganized biogeography-based-advancement (CBBO) [28], which will yield the worldwide ideal loads for tuning the profound learning system. In this manner, the enormous information grouping utilizing the proposed Rider Chaotic Biography improvement (RCBO) - based profound learning system in the flash engineering will give the ideal information arrangement. Figure 1 demonstrates the stream outline of proposed enormous information arrangement utilizing the proposed RCBO-based profound learning in sparkle design.

The execution of proposed enormous information order conspire in sparkle engineering will be done in JAVA and the dataset used for the examination will be the UCI AI store [26].

The proposed RCBO-based profound learning for huge information grouping is assessed by utilizing the measurements, for example, exactness and shakers coefficient that will be contrasted and the current works [2], [4], and [9].

Figure 1. Flow diagram of proposed big data classification



Framework

C. *Applications*

The uses of huge information order are shown as pursues,

1) Banking

The way toward overseeing and assessing the information of banks and other monetary administrations associations contains colossal measure of customer information which incorporates individual and security data. In banking and account, reference information, exchange and market information, exchange information can be organized or unstructured dependent on the gathered data. Subsequently, the grouping of enormous information empowers to deal with every one of the information in one spot in an organized way.

2) Cloud Computing

The correspondence between the servers utilizing data innovation produces enormous measure of information. These information needs to handled and put away for appropriate working. In this manner, the cloud is utilized as an online stockpiling model for preparing gigantic measure of information.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

3) Healthcare

The huge information contains gigantic measure of information that is accessible for human services suppliers to screen the wellbeing dangers. Therefore, the medicinal services data and the rising consideration for wellbeing has adjusted a major space in settling on key business choices.

4) Data mining

The huge information arrangement utilizes information digging for anticipating the important information as opposed to taking superfluous information. It uses the information and examination for distinguishing the prescribed procedures for the arrangement.

5) Stocks

It is more straightforward to dissect the drifting stocks according to the arranged outcome. The characterization can check the enthusiasm of the individuals moving toward the stocks.

V.CONCLUSION

In this paper, we have introduced a novel methodology for Big Data Classification dependent on Deep learning and Apache Spark engineering. We have made a thorough writing audit and discover the exploration holes in existing philosophies. In light of difficulties and research holes we casing out the examination goals. Our proposed methodology depends on two advancement calculation in particular ROA(Rider Optimization Algorithm) and (CBBO) Chaotic-Bio geology based-Optimization. We utilize the cross breed technique which is really another strategy for enhancement we named it as Rider Chaotic Biography improvement (RCBO) calculation. In future work we will execute the framework utilizing RCBO calculation for any of the application notice in this paper for Big information order. This framework will give the superior and best exactness for characterization.

ACKNOWLEDGMENT

The entire session of this survey paper work completion was a great experience providing me with great insight and innovation into learning various concepts of data science, big data, deep learning.

As is rightly said ,for the successful completion of any work people are the most important asset. This paper wouldnt be materialized without the cooperation of many of the people involved.

First and foremost, I am thankful to my supervisor Dr. B. Chaitanya Krishna for his leading guidane and sincere efforts in finalizing the topic and work. He took a deep intrest in correcting the minor mistakes and guided me through my journey so far.

REFERENCES

1. S. Ramírez-Gallego, B. Krawczyk, S. García, M. Woźniak, J. M. Benítez and F. Herrera, "Nearest Neighbor Classification for High-Speed Big Data Streams Using Spark," in IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 47, no. 10, pp. 2727-2739, October 2017.
2. M. Duan, K. Li, X. Liao and K. Li, "A Parallel Multi classification Algorithm for Big Data Using an Extreme Learning Machine," in IEEE Transactions on Neural Networks and Learning Systems, vol. 29, no. 6, pp. 2337-2351, June 2018.
3. Elsebakhi, E., Lee, F., Schendel, E., Haque, A., Kathireason, N., Pathare, T., Syed, N. and Al-Ali, R., "Large-scale machine learning based on functional networks for biomedical big data with high performance computing platforms," Journal of Computational Science, vol. 11, pp. 69-81, 2015.
4. W. Lin, Z. Wu, L. Lin, A. Wen and J. Li, "An Ensemble Random Forest Algorithm for Insurance Big Data Analysis," in IEEE Access, vol. 5, pp. 16568-16575, 2017.
5. Hernández, Á.B., Perez, M.S., Gupta, S. and Muntés-Mulero, V., Using machine learning to optimize parallelism in big data applications. Future Generation Computer Systems, vol. 86, pp.1076-1092, 2018.
6. Ramírez-Gallego, S., García, S., Benítez, J.M. and Herrera, F., "A distributed evolutionary multivariate discretizer for big data processing on apache spark," Swarm and Evolutionary Computation, vol. 38, pp. 240-250, 2018.
7. Karim, M.R., Cochez, M., Beyan, O.D., Ahmed, C.F. and Decker, S., "Mining maximal frequent patterns in transactional databases and dynamic data streams: a spark-based approach," Information Sciences, vol. 432, pp.278-300, 2018.
8. Salloum, S., Dautov, R., Chen, X., Peng, P.X. and Huang, J.Z., "Big data analytics on Apache Spark," International Journal of Data Science and Analytics, vol. 1, pp.145-164, 2016.
9. B. Zhao, H. Zhou, G. Li and Y. Huang, "ZenLDA: Large-scale topic model training on distributed data-parallel platform," in Big Data Mining and Analytics, vol. 1, no. 1, pp. 57-74, March 2018.
10. J. Yan, Y. Meng, L. Lu and C. Guo, "Big-data-driven based intelligent prognostics scheme in industry 4.0 environment," 2017 Prognostics and System Health Management Conference (PHM-Harbin), Harbin, pp. 1-5, 2017.
11. K. Zhang, Y. Tanimura, H. Nakada and H. Ogawa, "Understanding and improving disk-based intermediate data caching in Spark," 2017 IEEE International Conference on Big Data (Big Data), Boston, MA, pp. 2508-2517, 2017.
12. S. Caíno-Lores, J. Carretero, B. Nicolae, O. Yildiz and T. Peterka, "Spark-DIY: A Framework for Interoperable Spark Operations with High Performance Block-Based Data Models," 2018 IEEE/ACM 5th International Conference on Big Data Computing Applications and Technologies (BDCAT), Zurich, pp. 1-10, 2018.
13. G. Ditzler, S. Hariri and A. Akoglu, "High Performance Machine Learning (HPML) Framework to Support DDDAS Decision Support Systems: Design Overview," 2017 IEEE 2nd International Workshops on Foundations and Applications of Self* Systems (FAS*W), Tucson, AZ, pp. 360-362, 2017.
14. S. Ekanayake, S. Kamburugamuve, P. Wickramasinghe and G. C. Fox, "Java thread and process performance for parallel machine learning on multicore HPC clusters," 2016 IEEE International Conference on Big Data (Big Data), Washington, DC, pp. 347-354, 2016.
15. J. Fu, J. Sun and K. Wang, "SPARK – A Big Data Processing Platform for Machine Learning," 2016 International Conference on Industrial Informatics - Computing Technology, Intelligent Technology, Industrial Information Integration (ICIICII), Wuhan, pp. 48-51, 2016.
16. A. Gupta, H. K. Thakur, R. Shrivastava, P. Kumar and S. Nag, "A Big Data Analysis Framework Using Apache Spark and Deep Learning," 2017 IEEE International Conference on Data Mining Workshops (ICDMW), New Orleans, LA, pp. 9-16, 2017.
17. A. T. Hadgu, A. Nigam and E. Diaz-Aviles, "Large-scale learning with AdaGrad on Spark," 2015 IEEE International Conference on Big Data (Big Data), Santa Clara, CA, pp. 2828-2830, 2015.
18. Z. Han and Y. Zhang, "Spark: A Big Data Processing Platform Based on Memory Computing," 2015 Seventh International Symposium on Parallel Architectures, Algorithms and Programming (PAAP), Nanjing, pp. 172-176, 2015.
19. K. Kato, A. Takefusa, H. Nakada and M. Oguchi, "Consideration of parallel data processing over an apache spark cluster," 2017 IEEE International Conference on Big Data (Big Data), Boston, MA, pp. 4757-4759, 2017.
20. A. Koliopoulos, P. Yiapanis, F. Tekiner, G. Nenadic and J. Keane, "A Parallel Distributed Weka Framework for Big Data Mining Using Spark," 2015 IEEE International Congress on Big Data, New York, NY, pp. 9-16, 2015.

21. S. N. Lighari and D. M. A. Hussain, "Testing of algorithms for anomaly detection in Big data using apache spark," 2017 9th International Conference on Computational Intelligence and Communication Networks (CICN), Girne, pp. 97-100, 2017.
22. J. Lv, B. Wu, C. Liu and X. Gut, "PF-Face: A Parallel Framework for Face Classification and Search from Massive Videos Based on Spark," 2018 IEEE Fourth International Conference on Multimedia Big Data (BigMM), Xi'an, pp. 1-7, 2018.
23. M. A. Rahman, J. Hossen and V. C, "SMBSP: A Self-Tuning Approach using Machine Learning to Improve Performance of Spark in Big Data Processing," 2018 7th International Conference on Computer and Communication Engineering (ICCCE), Kuala Lumpur, pp. 274-279, 2018.
24. A. Sheshasaayee and J. V. N. Lakshmi, "An insight into tree based machine learning techniques for big data analytics using Apache Spark," 2017 International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT), Kannur, pp. 1740-1743, 2017.
25. S. Srivastava, A. Nigam and R. Kumari, "Work-in-Progress: Towards Efficient and Scalable Big Data Analytics: Mapreduce vs. RDD's," 2017 International Conference on Information Technology (ICIT), Bhubaneswar, pp. 272-275, 2017.
26. UCI machine learning repository, "https://archive.ics.uci.edu/ml/index.php", Acessed on February 2019.
27. D. Binu ; B. S Kariyappa, " RideNN: A New Rider Optimization Algorithm-Based Neural Network for Fault Diagnosis in Analog Circuits", IEEE Transactions on Instrumentation and Measurement, vol.68 , no.1 , pp.2 - 26, Jan. 2019.
28. Jie-Sheng Wang, and Jiang-Di Song , "Chaotic Biogeography-based Optimisation (CBBO) algorithm", IAENG International Journal of Computer Science, vol. 44, no. 2, 24 May 2017.
29. J. Dean and S. Ghemawat, "Mapreduce: Simplified data processing on large clusters," in *Proc. OSDI*, San Francisco, CA, USA, 2004, pp. 137–150.
30. A. Fernández *et al.*, "Big data with cloud computing: An insight on the computing environment, mapreduce, and programming frameworks," *Wiley Interdiscipl. Rev. Data Min. Knowl. Disc.*, vol. 4, no. 5, pp. 380–409, 2014.
31. T. White, *Hadoop, the Definitive Guide*. Sebastopol, CA, USA: O'Reilly Media, 2012.
32. Apache Hadoop Project. (2017). *Apache Hadoop*. [Online]. Accessed on Jan. 2017. [Online]. Available: http://hadoop.apache.org/
33. J. Lin, "Mapreduce is good enough? If all you have is a hammer, throw away everything that's not a nail!" *Big Data*, vol. 1, no. 1, pp. 28–37, 2012.
34. X. Meng *et al.*, "Mllib: Machine learning in apache spark," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 1235–1241, 2016.
35. A. Spark. *Machine Learning Library (MLlib) for Spark*. Accessed on Jan. 2017. [Online]. Available: http://spark.apache.org/docs/latest/mllib-guide.html.

AUTHORS PROFILE

Mr. Anilkumar Vishwanath Brahmane, Research Scholar, Department of Computer Science and Engineering, K L Deemed to be University, Vijaywada, Andhra Pradesh Pin – 520 002
His area of intrest is Big Data, Parallel Processing, Machine Learning.

Dr. B. Chaitanya Krishna, Professor, Department of Computer Science and Engineering, K L Deemed to be University, Vijaywada, Andhra Pradesh Pin – 520 002 Area of intrest is Computer Engineering, Data mining , Big Data, Machine Learning.