International Journal of Computer Trends and Technology
ISSN: 2231 – 2803 / https://doi.org/10.14445/22312803/IJCTT-V71I2P105

Volume 71 Issue 2, 25-31, February 2023

Review Article

Literature Review: An Efficient Clustering Approach to Big Data

Satish S. Banait[1], Tanuja B. Kaklij[2], Gauri K. Bankar[3], Srushti B. Hire[4] and Digvijay B. Wagh[5]

[1,2,3,4,5]*Department of Computer Engineering, K. K. Wagh Institute of Engineering Education & Research, Nashik (SPPU), Maharashtra, India*

Received: 24 December 2022 Revised: 25 January 2023 Accepted: 05 February 2023 Published: 17 February 2023

Abstract *- In today's era, data generated by scientific applications and the corporate environment has grown rapidly, not only in size but also in variety. There is difficulty in collecting, storing, transforming, and analyzing such big data. One of the major issues with big data is that the time taken to execute the traditional algorithms is larger, and it is very difficult to process a huge amount of data. Clustering is one of the popular data mining tasks. It is used in various domains. Machine learning is well-known for its unsupervised learning methods, such as the K-Means clustering algorithm. It has the benefits of easy implementation, good effect, and simplicity of the concept. But as the Internet expanded rapidly, the number of data collection points also increased, leading to the era of big data and information explosion. This research work proposes the IK-ABC (Improved K-Means - Artificial Bee Colony) Algorithm to address the issue of k-means clustering algorithms, such as low global search ability, sensitive selection of cluster center, initialization randomness, early development, and slow convergence of the original artificial bee colony Algorithm. A fitness function adapted to the K-means clustering method and a position update formula based on global guidance was created with MapReduce to speed up computation and increase the effectiveness of the iterative optimization process.*

Keywords *- Big Data, Clustering Algorithms, MapReduce, Swarm Optimization Techniques*.

1. Introduction

The field of data analysis and big data processing has seen a significant increase in the amount of huge data being generated and stored in recent years. Some studies argue that handling and using this huge data could become a new pillar of economics, scientific research, experimentation, and simulation. Indeed numerous chances of big data appearing in different areas similar to health (Enhancing the effectiveness of some treatments), transportation (reducing costs), finance (minimizing pitfalls), administration (decision stuff with high effectiveness and speed), social media, and government services.

However, in today's era, big data is also fraught with problems and has some quality issues like issues of scale, heterogeneity, privacy, timeliness, and visualization, at all stages of the analysis pipeline from data acquisition to result in interpretation. To improve data processing's effectiveness and usefulness, the most recent techniques and technologies are used to deal with this large data [1]. Another crucial data analysis technique is cluster analysis, which aims to categorize physical or abstract sets into related object classes so that items within the same group share a high degree of similarity and differ significantly from one another. There are different clustering algorithms used to manage large sets of data. But no clustering algorithm can solve all the Big Data issues [2]. Among them, the K-means algorithm is widely used because of its simplicity, but how to make it more compatible with the development of the era of big data still faces very big challenges like how to reduce the time complexity of the K-means algorithm and improve our clustering effect still needs further optimization [3].

In this research work, we propose K-Means Clustering Algorithm with Artificial Bee Colony (ABC) algorithm and MapReduce Framework. It is a powerful approach for solving large-scale clustering problems.

The contents of This research work are organized as follows: Section 2 reviews related work. Section 3 describes our proposed method of Improving the K Means Clustering Algorithm with the Artificial Bee Colony Algorithm (IK-ABC). Section 4 presents our experimental results and relevant performance analysis in terms of execution time and classification error. Finally, Section 5 presents our conclusions and discusses future direction.

2. Literature Survey

This section covers a broad review of big data difficulties, clustering algorithms, in particular, the K-Means Clustering Algorithm, the Artificial Bee Colony Algorithm, and the MapReduce Framework and big data applications.

The development of big data has led to the analysis of a wide variety of data formats, most of which are streaming in nature. As a result, conventional techniques have a difficult time meeting Big Data needs.

Big data are generated through internal and external sources of data; thus, existing systems fail to handle the unprecedented data. High-performance, highly scalable systems with advanced techniques are required to process valuable information. The study shows that the current tool and technology must be updated with time as the data is continuously growing [4]. The term "big data" describes a collection of numerical data generated by applying new technologies for either personal or professional usage. Big data analytics is used to analyze large amounts of data to find hidden patterns. The complexity of the analysis of this data, however, varied depending on the process that was needed [5], from traditional data analysis to the more current big data analysis and data analytics. The KDD process serves as the study's framework from a systems perspective. The unresolved problems with computing are discussed, resulting in quality, security, and privacy [6].

By grouping data using a variety of clustering algorithms, we set out to identify the day of the year with the greatest heart rate. A more effective clustering technique with improved accuracy, recall, and F-measure is produced via hybrid methodology. The hybrid technique produces the most clusters and includes each data point in each cluster [32]. EM and FCM clustering algorithms exhibit good performance in terms of the quality of the clustering outputs. Future research should address each clustering algorithm's shortcomings because none performs well for all evaluation criteria [8].

K-means clustering is a highly traditional clustering algorithm, and its use will increase over time. Future research may enhance the capability to handle large or multidimensional data sets. An area of study is the clustering of exponential data using K-Means [9]. A popular clustering method that is frequently used for clustering massive amounts of data is K-means. An effective method for clustering data points is presented in this research. The suggested approach guarantees that clustering is completed in O(nk) time [10]. However, K-means requires initial data point selection and nearest cluster assignment. This study explains how to more accurately assign data points to their nearest clusters and determine initial centroids using improved methodologies [11].

An analysis of previous work on artificial bee colony algorithm (ABC), ABC variations, and data clustering applications. ABC is a straightforward and adaptable method that requires less parameter tuning than other algorithms. The efficiency, precision, and usefulness of ABC in solving various optimization issues are demonstrated by numerous tests conducted in the pertinent literature [12]. ABC works on position updating formula and objective function. The iterative optimization procedure is more effective by using a position update formula based on local better and global best [13].

An artificial bee colony algorithm based on information learning (ILABC) could be useful for data structuring and data probation. The design of wireless telecommunications networks and the flow scheduling problem illustrate difficult optimization problems that can be solved with ILABC. Applying ILABC to more difficult issues may be worthwhile [14]. Our dataset's size has constantly been growing, making it challenging to cluster the data using conventional clustering algorithms. The fastest execution time is provided by the ABC system, which is also more effective for all sorts of data. To discover the optimal fitness value, the mapper phase simulates the behavior of an employed bee. In the reducer mode, the behavior of an observer bee is simulated to optimize the clusters [15].

The ease of use and quick convergence, the clustering algorithm has become a popular technique for cluster analysis. The IABC algorithm is suggested to solve the issues with the K-means clustering algorithm's randomly chosen initial centre points and poor global search capability [16]. The k-means algorithm challenges selecting an appropriate set of parameters, such as the number of clusters k and initial centroids. For the ABC algorithm, they have not discovered any attempts to date. A novel method to generate variable-length food sources for the ABC algorithm with a variable length (ABCVL) to supply the system with an appropriate level of diversity [17]. The ABC-based cluster has improved the influence of the initial center value and increased inter-group variation and similarity in the clustering [18]. A hybrid clustering algorithm based on modified ABC and K-Means algorithms. The relative fitness of each person - the ratio between their individual and overall fitness is used to create a roulette wheel. In the onlooker bee phase, variable tournament selection is used instead of roulette wheel selection [33].

This study aimed to provide an overview of the MapReduce ideas used in big data analytics. To analyze large data, which is unstructured data like web data, Google developed Map Reduce [20]. Big data and related technologies can positively impact the company's operations. A few guidelines must be followed to acquire fast and beneficial results from big data. Programming MapReduce using the Hadoop framework, which is an open-source system, accelerates the processing of massive amounts of data [21]. Without any prior programming knowledge, programmers can simply grasp the MapReduce framework. Load balancing, fault tolerance, serialization, and parallelization are no longer required [22]. The data mining environment of the Hadoop cluster is used to study the K-means method. With the help of the improved algorithm, catering decision-makers may identify high-value consumer segments and provide superior service. The k-Means algorithm for processing data mining has superior expansion performance and mining efficiency in a cluster of cloud computing platforms, which has been demonstrated [23].

The K-Means Clustering Algorithm offers a reliable and effective method for classifying data that have similar

features. It lowers the implementation costs associated with handling such massive data volumes via a distributed network. Reducing the number of iterations needed to finish a task allows for improvements [2]. A parallel K-means method based on Hadoop is given in work with quite good findings for data processing effectiveness and convergence. As the amount of data increases, the acceleration effect is better for processing huge amounts of data, especially in the MapReduce architecture [25]. The standard K-means method has been enhanced. The problem of the K-Means initial center point sensitivity was resolved by the modified approach, which successfully identified the initial clustering centers. Large data processing was made possible by better algorithm parallelization. The performance of the K-means algorithm has been increased, and both techniques significantly improve results [26]. K-means algorithm improves MapReduce design using an iteration-saving technique. They illustrate that this keeps 80% of the clustering accuracy while reducing the number of iterations and execution time in clustering techniques [2].

An effective artificial bee colony for MapReduce-based large-scale data clustering is developed. In the Hadoop system, the ABC could be used to streamline the clustering of enormous amounts of data. It provides an adequate level of grouping and performance in comparison to more current methods [28]. The novel optimization method has effective search capabilities in the solution space, and a pattern is applied to achieve the best outcomes with fewer iterations. Many methods enhance the search quality and fast local search time in global search by integrating and extracting the features of both MapReduce and a specific method [29]. MapReduce's parallelization capabilities make using the Artificial Bee Colony technique simple. Each member of the population just needs to look in a very small area, which allows them to find the answer more quickly. Because the particles continually update themselves after each iteration, the proposed model for parallel ABC can use a huge population but cannot be used with a large dataset [30]. The Modified Artificial Bee Colony Algorithm is the optimization algorithm we used (MABC). A method for utilizing the map-reducing algorithm to solve resource issues in clouds. With the aid of the optimization algorithm, the MapReduce algorithm creates a further improved solution. The suggested approach to resource problem reduction works better because it requires less space for data storage [31].

3. Proposed Approach

This section demonstrates the details of the proposed approach, as shown in Figure 1. The proposed approach used three fundamental parts to get the best model as a result. Clustering algorithms, such as Artificial Bee Colony (ABC) and K-Means, are powerful tools for analyzing and understanding large datasets. However, their use can be hindered by the computational challenges involved in processing large amounts of data. MapReduce is a distributed computing framework that allows for the efficient processing of large datasets by breaking them down into smaller, more manageable chunks.

The combination of ABC, K-Means, and MapReduce has been the subject of recent research, as it has the potential to improve the performance and scalability of clustering algorithms for big data processing. This literature review paper aims to provide an overview of the current state of research on using ABC and K-Means algorithms with MapReduce for big data clustering and to identify the key challenges and opportunities in this area.

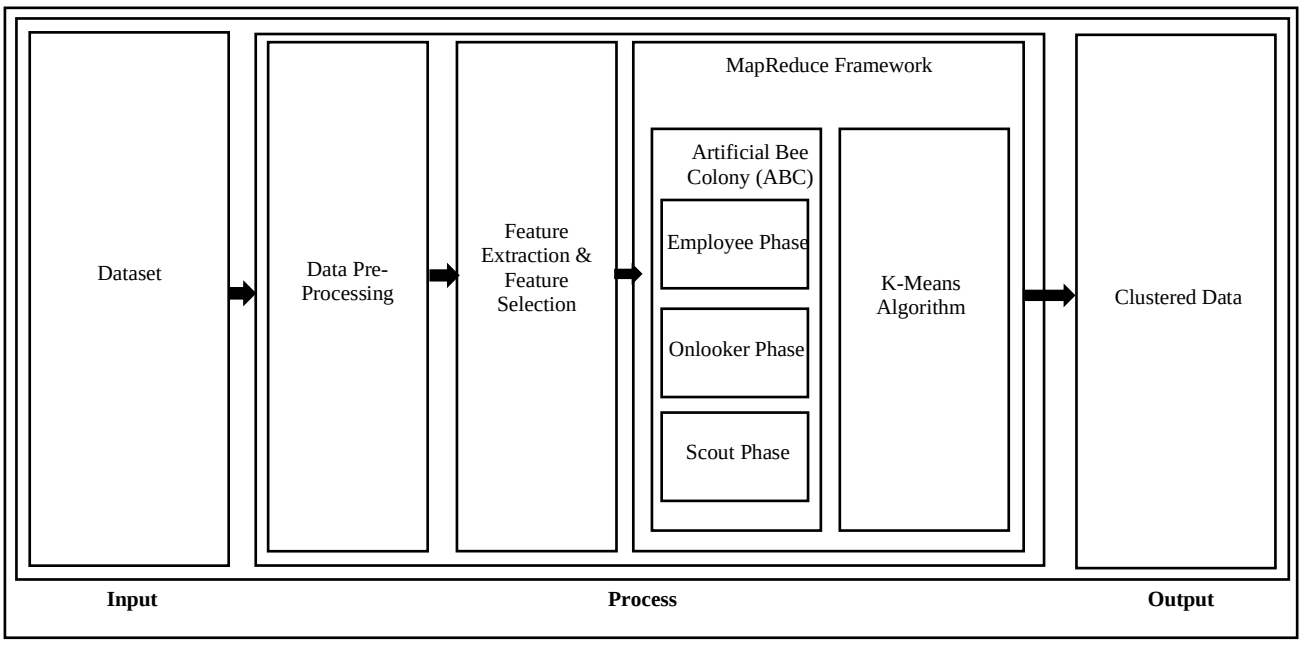

Fig. 1 Block Diagram

3.1. K-Means Clustering Algorithm

The K-Means Clustering is an unsupervised learning algorithm used to group similar data points together. The k-means clustering technique divides a dataset into a predetermined class number of k based on minimizing the error. The center of the cluster *Ej* (j = 1, 2,⋯, k) is used to represent the cluster and distance is used to measure similarity. *D(xi, xj)* denotes the Euclidean distance between two data items, xi and xj, and its calculation method is as follows

$$D(x_i, x_j) = \sqrt{(x_{i1} - x_{i2})^2 + \cdots (x_{iL} - x_{iL})} \quad (1)$$

where *L* denotes the number of data object attributes.

The algorithm's objective is to reduce the within-cluster variance, which is a gauge of how similar the data points are to one another within a cluster. Error square and SSE are employed as objective functions to assess the clustering effectiveness and depict how tightly the samples are clustered. The higher the sample similarity, the smaller the SSE is. The following is the SSE calculating formula:

$$SSE = \sum_{j=1}^{k} \sum_{x \in E_j} D(x, e_j) \quad (2)$$

$$e_j = \frac{1}{n_j} \sum_{x \in E_j} x \quad (3)$$

where *nj* denotes the number of sample data in the j^{th} cluster E_j.

3.2. Artificial Bee Colony Algorithm

Use The artificial bee colony algorithm is a swarm intelligence algorithm inspired by the foraging behavior of honeybees. It simulates the foraging behavior of honeybees by creating a population of artificial bees, each representing a candidate solution. The bees search for nectar sources which correspond to the optimal solutions in the problem space.

The algorithm comprises three components: a leader, a follower, and a scouter. The leader identifies a specific food source and conveys details about that source. In the dance region of the hive, the follower bees first waited for the leader bees to share information about the food source, after which they selected one and started further exploration around it. Scouters are in charge of randomly searching for new food sources.

The fundamental artificial bee colony algorithm can be divided into four stages.

3.2.1. Initialization Phase

The algorithm begins by randomly initializing the bee population. Each N food source represents a feasible solution created randomly in the feasible solution space. The precise formula is as follows:

$$x_{i,j} = x_j^{min} + random(0,1) \times (x_j^{max} - x_j^{min}) \quad (4)$$

where D is the dimension of the feasible solution and i=1, 2,⋯, N; j=1, 2,⋯, D. The j^{th} parameter's upper and lower bounds are represented by the variables x_j^{max} and x_i^{max}. Set a counter as well, with a value of 0, for each food source.

3.2.2. Leader Search Phase

The leader locates a new food source v_i in the neighborhood of the corresponding food source with the help of the following formula:

$$v_{i,j} = x_{i,j} + (-1 + 2 \times random) \times (x_{i,j} - x_{k,j}) \quad (5)$$

where k is a set of randomly chosen food sources distinct from i, k ≠ i, the "greedy selection" method is used for the new and old food sources xi and vi, i.e., the new and old food source quality is compared, if the quality of the new food source is higher, it is kept, and its counter is set to 0. Otherwise, it is removed. If necessary, keep the old food source, but add one to its counter.

3.2.3. Follower Search Phase

Follower bees choose a food source from the available food source by spinning a roulette wheel. The likelihood of choosing each food source is as follows:

$$P_i = \frac{fit_i}{\sum_{j=1}^{N} fit_j} \quad (6)$$

where fit_i is a measure of the quality of the food source and is determined using the formula below:

$$fit_i = \begin{cases} \frac{1}{1+f_i} & f_i \geq 0 \\ 1 + |f_i| & otherwise \end{cases} \quad (7)$$

where f_i is the value of the objective function.

The follower subsequently conducted operations associated with greedy selection after choosing the food source and searching the field in accordance with phase 2.

3.2.4. Scouter Search Phase

A specific scout bee search mode was introduced to the bee colony algorithm to prevent the loss of population diversity throughout the evolution phase. When a food source's counter value exceeds a predetermined threshold limit, the food source can be considered exhausted or abandoned, and the corresponding lead bee changes roles to become the scout bee. A new food source is then generated randomly using the phase 1 method in the feasible solution space.

3.3. Improved K-Means – Artificial Bee Colony (IK-ABC) Algorithm

The Improved K-Means Clustering Algorithm with Artificial Bee Colony (ABC) algorithm and MapReduce is a powerful approach for solving large-scale clustering problems. A popular technique for grouping related objects in data is the K-Means algorithm. However, the traditional

K-Means algorithm may suffer from poor convergence or get stuck in local optima. The Artificial Bee Colony (ABC) algorithm is a population-based optimization algorithm inspired by the foraging behavior of honeybees, which can be used to overcome these limitations of K-Means.

The ABC algorithm is used to improve the initialization of the centroids in the K-Means algorithm, which can improve the quality of the final clusters. However, the fundamental artificial bee colony algorithm has the following two drawbacks: 1) Low efficiency is caused by random initialization 2) slow convergence is caused by one-dimensional domain search. In this research, the maximum and minimum distance product method is used to initialize the sealed group to overcome its randomness. The new fitness function and the global guide factor's position change formula are used for iterative optimization.

3.3.1. Maximum and Minimum Distance Product

The ABC algorithm generates candidate solutions by randomly selecting the search space and then calculating the fitness function of each solution. The maximum and minimum distance product is used to adjust the search space and improve the solutions' quality. This method minimizes the sensitivity of the k-means algorithm to the initial point while simultaneously overcoming the randomness of colony initialization.

3.3.2. Fitness Function

Using a fitness function, the ABC algorithm can determine the quality of candidate solutions and focus its search efforts on solutions with a higher fitness value. This helps to ensure that the algorithm converges quickly to the optimal solution and provides a more efficient way to search the solution space. Fitness Function is represented by the following formula:

$$fitness_i = \frac{CM_i}{Dist_i}\ i = 1,2,\cdots,N \quad (8)$$

where CM_i is the number of points belonging to class i; $Dist_i$ is the sum of distances between all objects in class i and center

$$C_i, Dist_i = \sum_{x_j \in C_i} d(x_j, C_k), Dist = \sum_{j=1}^{k} \sum_{x_i \in C_j} d(x_i, C_j) \quad (9)$$

3.3.3. Position Updating Formula

Position Updating Formula is used to update the position of each food source, which represents a candidate solution to the optimization problem based on the information gathered by the employed and onlooker bees. The position updating formula considers the quality of the food source, as represented by its fitness function, and the direction towards a potentially better food source, as determined by the employed bee. The formula effectively balances exploitation (searching for the current best solution) and exploration (searching for new and potentially better solutions) to find an optimal solution.

$$v_{ij} = x_{ij} + r_{ij}(x_{mj} - x_{kj}) + \mu\,(x_{best,j} - x_{i,j}) \quad (10)$$

where v_{ij} is a new position generated near x_{ij}; k, m, and j are random numbers generated by random formulas, k, m $\in\{1,2, \cdots, N\}$, k and m are mutually exclusive, and neither is equal to i; $r_{ij} \in [-1,1]$; $\mu \in [0,1]$ is a random number; $x_{\text{best,j}}$ is the most abundant source of honey.

where v_{ij} is a newly produced position that is close to x_{ij}; k, m, and j are random integers generated using random formulae k, m $\in\{1,2,\cdots, N\}$, k and m are mutually exclusive; neither of them is equal to $r_{ij} \in [-1,1]$; $\mu \in [0,1]]$ is a random number; $x_{\text{best, j}}$ is the honey source with the greatest proportion.

Additionally, the MapReduce framework is utilized to perform distributed computing on the dataset, allowing for efficient processing of large amounts of data. The mapper function assigns each data point to the nearest centroid, and the reducer function updates the centroids based on the data points assigned to them.

The algorithm for combining the Artificial Bee Colony (ABC) optimization algorithm with the K-means clustering algorithm and MapReduce can be outlined as follows:

1) Initialize the population of bees, where each bee represents a candidate solution for the initial centroids of the K-means clustering algorithm.
2) Split the dataset into smaller subsets and distribute them to different worker nodes using the MapReduce framework.
3) Each worker node runs the K-means clustering algorithm using the bee's solution as the initial centroids for the subset of data it received. The quality of the clustering solution, such as the sum of squared errors, determines the fitness value.
4) Use the ABC algorithm to optimize the population of bees by performing the following steps:
   a. Employed bees: The employed bees update the solution of the bee they are associated with by using a neighbourhood search strategy.
   b. Onlooker bees: The onlooker bees select a solution to update based on the probability of the fitness of each solution.
   c. Scout bees: The scout bees search for new solutions by randomly generating new solutions and replacing the worst solutions in the population.
5) The worker nodes return their results to the master node, where they are combined to form a global population of bees.
6) Repeat steps 3 to 5 until the stopping criteria are met, such as a maximum number of iterations or a satisfactory level of convergence.
7) The final solution is the best bee, which represents the optimal initial centroids for the K-means clustering algorithm.

4. Conclusion

The review of prior research on big data, the K-means clustering, the Artificial Bee Colony Algorithm (ABC), and the MapReduce Framework is presented in this research work. The proposed approach provides an efficient clustering approach to big data. The IK-ABC (Improved K-Means & Artificial Bee Colony) Algorithm is the solution to the issues with the k-means clustering algorithm as well as with the ABC algorithm. The proposed IK-ABC approach increases convergence speed and reduces the computing time using the MapReduce framework while dealing with massive data. So that big data will be handled effectively and efficiently.

References

[1] Guma Abdulkhader Lakshen, Sanja Vranes, and Valentina Janev, "Big Data & Quality- A Literature Review," *24th Telecommunications forum TELFOR*, pp. 1-4, 2016. *Crossref*, https://doi.org/10.1109/TELFOR.2016.7818902

[2] Prajesh P. Anchalia, Anjan K. Koundinya, and Srinath N. K, "Map Reduce Design of K-means Clustering Algorithm," *IEEE International Conference on Information Science and Applications (ICISA)*, pp. 1-5, 2013. *Crossref*, https://doi.org/10.1109/ICISA.2013.6579448

[3] Chen Jie et al., "Review on the Research of K-means Clustering Algorithm in Big Data," *IEEE, International Conference on Electronics and Communication Engineering*, pp. 107-111, 2020. *Crossref*, https://doi.org/10.1109/ICECE51594.2020.9353036

[4] R Rawat and R Yadav, "Big Data: Big Data Analysis, Issues and Challenges and Technologies," *IOP Conference Series-Materials Science and Engineering*, vol. 1022, 2021. *Crossref*, https://doi.org/10.1088/1757-899X/1022/1/012014

[5] Abdulbaset S. Albaour, and Yousof A. Aburawe, "Big Data: Review Paper," *International Journal Of Advance Research And Innovative Ideas In Education*, vol. 7, no. 1, 2021.

[6] Chun-Wei Tsai et al., "Big Data Analytics: A Survey," *Journal of Big Data*, vol. 2, no. 20, 2015. *Crossref*, https://doi.org/10.1186/s40537-015-0030-3

[7] Fatema Jamnagarwala, and P.A.Tijare "Implementation of Data Mining With lustering of Big data for Shopping mall's data using SOM and K-means Algorithm," *International Journal of Computer Trends and Technology*, vol. 67, no. 12, pp. 3-7, 2019. *Crossref*, https://doi.org/10.14445/22312803/IJCTT-V67I12P102

[8] Adil Fahad et al., "A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis," *IEEE Transactions on Emerging Topics in Computing*, vol. 2, no. 3, pp. 267-279, 2013. *Crossref*, https://doi.org/10.1109/TETC.2014.2330519

[9] Bao Chong, "K-Means Clustering Algorithm: A Brief Review," *Academic Journal of Computing & Information Science*, vol. 4, no. 5, 2021. *Crossref*, https://doi.org/10.25236/AJCIS.2021.040506

[10] Shi Na, Liu Xumin, and Guan Yong "Research on k-means Clustering Algorithm", *3rd Intl Symposium on Intelligent Information Technology and Security Informatics*, pp. 63-67, 2010. *Crossref*, https://doi.org/10.1109/IITSI.2010.74

[11] Unnati R. Raval, and Chaita Jani, "Implementing & Improvisation of K-means Clustering Algorithm," *International Journal of Computer Science & Mobile Computing*, vol. 5, no. 5, pp. 191-203, 2016.

[12] Ajit Kumar, Dharmender Kumar, and S. K. Jarial, "A Review on Artificial Bee Colony Algorithms and Their Applications to Data Clustering," *Cybermetics and Information Technologies*, vol. 17, no. 3, pp. 3-28, 2017. *Crossref*, https://doi.org/10.1515/cait-2017-0027

[13] Yi Yang, and Ke Luo, "An Artificial Bee Colony Algorithm Based on Improved Search Strategy," *2nd International Conference on Artificial Intelligence and Information*, no. 191, pp. 1-4, 2021. *Crossref*, https://doi.org/10.1145/3469213.3470398

[14] Wei-Feng Gao et al., "Artificial Bee Colony Algorithm Based on Information Learning," *IEEE Transactions On Cybernetics*, vol. 45, no. 12, pp. 2827-2839, 2015. *Crossref*, https://doi.org/10.1109/TCYB.2014.2387067

[15] S. Sudhakar Ilango et al., "Optimization using Artificial Bee Colony Based Clustering Approach for Big Data," *Cluster Computing*, vol. 22, no. 5, pp. 12169-12177, 2019. *Crossref*, https://doi.org/10.1007/s10586-017-1571-3

[16] Zhenrong Zhang, Jiayi Lan and Zhenrong Zhang, "K-Means Clustering Algorithm Based on Bee Colony Strategy," *2nd Internation Conference on Signal Processing and Computer Science*, 2021. *Crossref*, https://doi.org/10.1088/1742-6596/2031/1/012058

[17] Sabreen Fawzi Raheem, and Maytham Alabbas "Optimal K-Means Clustering Using Artificial Bee colony Algorithm with Variable Food Sources Length," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 12, no. 5, 2022. *Crossref*, http://doi.org/10.11591/ijece.v12i5.pp5435-5443

[18] Ting-En Lee, Jao-Hong Cheng, and Lai-Lin Jiang, "A New Artificial Bee Colony Based Clustering Method & its Application to the Business Failure Prediction," *International Symposium on Computer, Consumer and Control*, pp. 72-75, 2012. *Crossref*, https://doi.org/10.1109/IS3C.2012.28

[19] Ranjit Rajak, Satish Chaurasiya, and Anjali Choudhary, "Integration of Big Data and Cloud Computing: Tools, Issues, and Reliability," *International Journal of Engineering Trends and Technology*, vol. 70, no. 11, pp. 170-177, 2022. *Crossref*, https://doi.org/10.14445/22315381/IJETT-V70I11P218

[20] P. Sudha, and R. Gunavathi, "A Survey Paper on Map Reduce in Big Data," *International Journal of Science and Research*, vol. 5, no. 9, 2016.

[21] Seema Maitreya, and C.K. Jha, "MapReduce: Simplified Data Analysis of Big Data," *Procedia Computer Science*, vol. 57, pp. 563-571, 2015. *Crossref*, https://doi.org/10.1016/j.procs.2015.07.392

[22] Muthu Dayalan, "MapReduce: Simplified Data Processing on Large Cluster," *International Journal of Research and Engineering,* vol. 5, no. 5, PP. 399-403, 2018. *Crossref,* https://doi.org/10.21276/ijre.2018.5.5.4

[23] Hongqin Wang et al., "Research & Application of Improved K-Means Based on MapReduce," *Journal of Physics: Conference Series*, vol. 1651, no. 1, pp. 12074, 2020. *Crossref,* https://doi.org/10.1088/1742-6596/1651/1/012074

[24] Oussama Lachiheb, Mohamed Salah Gouider, and Lamjed Ben Said, "An Improved MapReduce Design of Kmeans with Iteration Reducing for Clustering Stock Exchange the Very Large Datasets," *11th International Conference on Semantics, Knowledge and Grids*, pp. 252-255, 2015. *Crossref,* https://doi.org/10.1109/SKG.2015.24

[25] Jiyang Jia, Hui Xie, and Tao Xu, "Design and Implementation of K-Means Parallel Algorithm Based on Hadoop," *2nd International Conference on Artificial Intelligence and Information Systems*, no. 206, pp. 1-4, 2021. *Crossref,* https://doi.org/10.1145/3469213.3470413

[26] Li Ma et al., "An Improved K-means Algorithm based on Mapreduce and Grid," *International Journal of Grid and Distributed Computing,* vol. 8, no.1, pp.189-200, 2015. *Crossref,* https://doi.org/10.14257/ijgdc.2015.8.1.18

[27] K.Iswarya, "Security Issues Associated With Big Data in Cloud Computing," *SSRG International Journal of Computer Science and Engineering* , vol. 1, no. 8, pp. 1-5, 2014. *Crossref,* https://doi.org/10.14445/23488387/IJCSE-V1I8P101

[28] Anan Banharnsakun, "A MapReduce-Based Artificial Bee Colony for Large Scale Data Clustering," *Pattern Recognition Letters*, vol. 93, pp. 78-84, 2016. *Crossref,* https://doi.org/10.1016/j.patrec.2016.07.027

[29] Parikshit Patil et al., "Optimization of Data using Artificial Bee Colony Optimization with Map Reduce," *ITM Web of Conf*erence, vol. 32, no. 3031, pp. 1-6, 2020. *Crossref,* https://doi.org/10.1051/itmconf/20203203031

[30] Nupur Bansal, Sanjay Kumar, and Ashish Tripathi, "Application of Artificial Bee Colony Algorithm Using Hadoop," *IEEE 3rd International Conference on Computing for Sustainable Global Development,* pp. 3615-3619, 2016.

[31] S.A.Gowri Manohari, and S.Jawahar "Large Biological Dataset Analysis Using Enhanced Map Reducing Method With Modified Artificial Bee Colony Optimization (MABC)," *Journal of Emerging Technologies and Innovative Research,* vol. 5, no. 12, 2018.

[32] Satish S. Banait, S. S. Sane and Sopan A.Talekar, "An Efficient Clustering for Big Data Mining", *International Journal of Next-Generation Computing*, vol. 13, no. 3, pp. 702-717, 2022.

[33] Ajit Kumar, Dharmender Kumar, and S. K. Jarial, "A Novel Hybrid K-Means & Artificial Bee Colony Algorithm Approach for Data Clustering," *Decision Science Letters*, vol. 7, no. 1, pp. 65–76, 2018. *Crossref,* https://doi.org/10.5267/j.dsl.2017.4.003