International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
DOI: https://doi.org/10.17762/ijritcc.v11i10.8517
***Article Received:** 02 August 2023 **Revised:** 29 September 2023 **Accepted:** 13 October 2023*

An Enhanced Query Optimization Implemented in Hadoop using Bio-Inspired Algorithm with HDFS Technique

[1]Abhijit Banubakode, [2]Vandana C. Bagal, [3]Vishwanath S. Mahalle, [4]Maqsood A. Ansari, [5]Chhaya S. Gosavi, [6]Archana Chaugule

[1]MET Institute of Computer Science
Maharashtra, Mumbai, India
e-mail: abhijitsiu@gmail.com
[2]K. K.Wagh Institute of Engineering Education and Research, Nashik,Maharashtra, India
e-mail:vcbagal@kkwagh.edu.in
[3]Shri Sant Gajanan Maharaj College of Engineering
Shegaon, Maharashtra, India
e-mail: vvntvsm@gmail.com
[4]Smt. Kashibai Navale College of Engineering
Pune, Maharashtra,India
e-mail: maqans@gmail.com
[5]MKSSS's Cummins College of Engineering for Women
Pune,Maharashtra,india
e-mail: chhaya.gosavi@cumminscollege.in
[6]Pimpri Chinchwad College of Engineering and Research
Pune,Maharashtra,India
e-mail: archna.ajit27@gmail.com

Abstract: A more effective method for massive data query optimization using HDFS and the Bio-inspired algorithm. Big Data configuration and query optimization are the two phases of the process. To remove redundant data, the input data is first per-processed using HDFS. Then, utilizing entropy calculation, features like closed frequent pattern, support, and confidence are extracted and managed. The Bio-inspired Horse Herd approach is used to group pertinent information based on this outcome. In the second step, the Big Data queries are used to obtain the same features. The optimized query is then located using the Bio-inspired technique, and the similarity assessment procedure is run. The proposed algorithm, according to this research, outperforms other ones that is unique in use. It is challenging to determine the veracity of this claim without more information regarding the experimental setup and the precise measures employed to assess the algorithm's effectiveness. Furthermore, it is unknown how the proposed algorithm stacks up against other cutting-edge query optimization methods. Finally, the assess has efficiency of using this method, more optimistic query achieved and comparison analysis are proved.

Keywords: Hadoop Distributed File System (HDFS), Bio-inspired Optimization algorithm,Secure Hash Algorithm, Big Data and Query processing.

I. INTRODUCTION

The value of big data analysis in both professional and academic environments. For example, net organizations accumulate full-size volumes of statistics from a selection of resources, including provider logs, site crawlers, and click- streams. it's miles called "big data" due to the fact the volume of statistics being amassed exceeds the ability and processing pace of traditional storage structures. The requirement for software structures that can cope with dynamic, multi-goal optimization problems in big facts. A computing platform created in particular for processing massive data is known as a huge facts processing platform. Currently, databases and business practices studies is extra worried with performance than energy efficiency. The article does acknowledge, but, that big data can't be treated by means of conventional relational database management structures (RDBMS) or conventional statistical strategies [1, 2].

To address such large statistics processing, many organizations depend on particularly disbursed software program structures running on vast clusters of commodity equipment. As a way to examine huge statistics sets, query

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
DOI: https://doi.org/10.17762/ijritcc.v11i10.8517
***Article Received:** 02 August 2023 **Revised:** 29 September 2023 **Accepted:** 13 October 2023*

optimization is critical considering that it can reduce the quantity of queries needed to examine the statistics. Given the developing significance of statistics in a spread of disciplines, consisting of the development of consumer bases, research, and marketplace analysis, the passage factors [3]. Numerous query plans for finishing queries are analyzed in many relational database management systems, and a terrific query plan is discovered to limit using precise resources, inclusive of I/O. To procedure data extra efficiently, an appropriate query get right of entry to method may be chosen with using this optimization [4].

Big data analytics are regularly accomplished by means of developing and running queries in sq.-like languages that are supported with the aid of structures like Hadoop, Scope, and Spark. An open-supply framework known as Hadoop makes it viable to broaden massive-scale, records-extensive computing programs. It makes use of algorithms which are based on Google's in-intensity examine and information in processing extensive amounts of statistics [5]. The Map-lessen paradigm, which enables disbursed processing of big datasets over clusters of computers, is a part of the structure this is supposed to make massive-scale batch processing less difficult to apply. Big statistics units are also saved the use of the Hadoop dispensed document system [6].

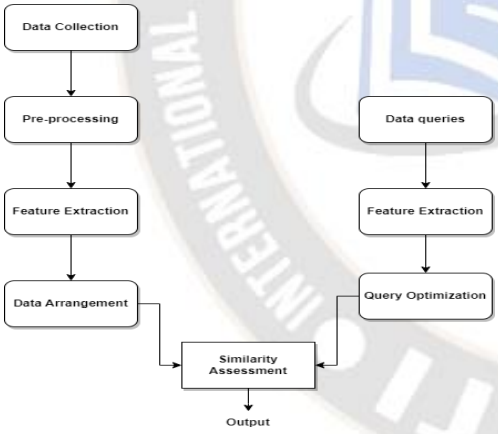

Figure 1. Traditional Block Diagram of Query Optimization based Framework for Big Data categorization and Pre-processing [7]

In view that Google first announced MR in 2004, it has grown to be a properly-liked framework for managing big datasets. Primarily based at the MR paradigm, the modern-day open-source Hadoop challenge gives the parallel processing of widespread volumes of records, automatic facts splitting, records distribution, fault tolerance, & load balancing control, main to dependable in parallel throughout a dispensed network of dozens to heaps of commodity machines via breaking them up into smaller, greater conceivable chunks. This approach permits for amazing scalability, availability, and fault tolerance because the machine can preserve jogging even if a few machines malfunction [8].

The problems of massive-scale facts analysis, which can also include processing tens or hundreds of terabytes of facts. The most broadly used open-source processing engine for BD is Apache Spark, which has an extensive variety of libraries and rich language-integrated APIs [9]. The primary components of the Spark API are collections of Java/Python gadgets, where users can execute arbitrary Python or Java operations the usage of operators like map and group.

The draft framework for this take a look at is systematized as section 2 surveys the related research on the advised approach. Section 3 prefers a succinct review of the proposed method, segment four explores the investigational final results, and section 5 deduces the paper.

II. LITERATURE SURVEY

Data locality and Hadoop [10] parameter tuning in dependable and uniform cluster environments make up the majority of the related attempt for enhancing Map reduce performance. [11] Cautioned have been used to categorize this system. The experimental findings confirmed that the use of iHOME on Hive, the overall execution time of 8 join queries was reduced. However, this machine nonetheless required some refinements.

A heuristic-centric method changed into proposed in [12] as a option to the Multi be a part of question Ordering (or MJQO) problem. This algorithm blended "2" critical seek algorithms: cuckoo and tabu seek. The simulation found out some exciting outcomes about the proposed algorithm and recommended that it could solve the MJQO trouble quicker than the present methods. A framework for growing and constructing green databases changed into furnished in [13]. First, a method for modelling the power cost of question plans for the duration of query processing based on their styles of aid intake changed into proposed. The basics of plan evaluation had been then examined. The evaluation method made use of the trade-offs among electricity and performance of plans, using the price model as a basis. The results of the experiments showed that this framework might also considerably lessen electricity use and boost power performance the usage of specific and trustworthy statistical records [14].

so as to reduce general value, [15] intensity category was created for PGNN searches, in which candidate object possibilities had been mechanically chosen based on diverse queries [16]. The PGNN query become then chosen to have

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
DOI: https://doi.org/10.17762/ijritcc.v11i10.8517
Article Received: 02 August 2023 Revised: 29 September 2023 Accepted: 13 October 2023

a few beneficial features. The extreme gaining knowledge of gadget (ELM)-based DCA (intensity type algorithm), which used a plurality balloting method, turned into then supplied. The effects showed that the ELM classifier's intensity-set prices are decrease general than the default values.

Added optimization procedures for recurrent queries in BD analysis in [17]. The MR regular window slice algorithm was implemented on this method to reuse recurrent queries. Moreover, the redundant facts become eliminated at the same time as input records was being loaded the use of exceptional-grain scheduling. Next, it evolved the MR late scheduling approach for facts scheduling, which greater information processing and optimized the scheduling of computation sources in the MR cluster. The results of the trial on diverse workloads verified that the algorithms exceeded the satisfactory techniques.

[18] Provided an approach as an addition approach became discovered to discover better answers, kind of 91.five-122.four% quicker than formerly a hit methods, in line with experimental opinions of this technique. This turned into completed by using disposing of the useless operations or even providing vital meta-facts as comments data to other marketers (ants).

In order to account for slow storage, it proposed the MOTH multi-query optimization device, which uses metadata and histograms had been evolved by using MOTH [19]. in step with the trial findings, the three absolutely reused-centric plans carried out better than the Naive technique plan on common with the aid of forty%, 45%, and 50%, respectively. On MR, the 2 partly reused-based totally plans fared higher on average than the Naive approach plan by 22% and 27%, respectively.

Our in advance studies focused on debunking common myths approximately large facts processing in dynamic and high quality conditions and supplying enormous insights into the underlying causes and potency of massive records processing. Our in advance studies pursuits to pave the way for future massive records processing optimization or avoidance that is the focus of this book.

III. PROPOSED METHODOLOGY

Companies keep many databases to save and system Big Data (BD), that's extraordinarily volumetric and has a ramification of statistics fashions. Business achievement depends on querying and BD analysis for perception. The use of each the HDFS map-reduce technique and the bio-inspired set of rules, this paintings stepped forward the question optimization procedure in BD.

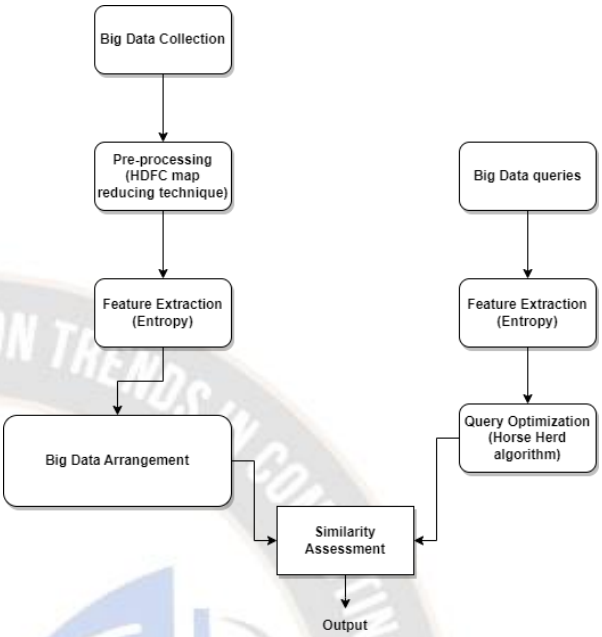

Figure 2. Proposed Block diagram

- BigData business enterprise calls for a methodical and effective approach to prepare and keep huge quantities of facts. The following are some famous strategies for organizing massive statistics:
- Distributed storage: using an allotted report device, along with Hadoop disbursed document gadget or Apache Cassandra, is a famous approach. big facts volumes can now be stored and accessed more effortlessly thanks to these systems' help for facts distribution across numerous nodes.
- Data partitioning: Dividing records into smaller subsets is every other manner. Information may be divided, for instance, in step with time, area, or every other characteristic. The records processing and analysis are facilitated by using this.
- Compression: large information units can eat up a variety of garage space, consequently compression is a great concept. It could use compression techniques like gzip or bzip2 to limit the amount of garage required.
- Statistics indexing: Indexing can assist huge facts units' search and question instances. It is feasible to set up indexes on unique qualities to make it easier to look for precise records.
- Data normalization: Organizing information into a standardized format is known as normalization. This will facilitate statistics assessment and evaluation throughout many records units.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
DOI: https://doi.org/10.17762/ijritcc.v11i10.8517
***Article Received:** 02 August 2023 **Revised:** 29 September 2023 **Accepted:** 13 October 2023*

➢ Statistics archiving: finally, its miles viable to shop statistics this is no longer required for evaluation. As an end result, garage area is freed up and handling the facts that is nevertheless in use is made easier.

A. Pre-processing

The pre-processing of the input data was carried out during this phase. First, it uses the Secure Hash Algorithm (SHA-512) to determine the HV for each piece of data. Then, using HDFS, the MR process is carried out using the HV as its focal point. The subsections below provide an explanation of the SHA-512 and HDFS processes. In [20] for the SHA-512 method used to calculate the hash value of large amounts of data. Blocks of the augmented data are created. The following stage updates a 512-bit buffer.

$$fH(Dn) = \{H(D1), H(D2), H(D3), \ldots\ldots, H(Dn)\} \quad (1)$$

WhereH (Dn) stands for the HV of the n-number of data, and fH (Dn) stands for find the HV of the n-number of data.

B. HDFS

A distributed file system called HDFS is employed to safely and effectively store and handle huge amounts of data. It is made to function with the HadoopMapReduce (MR) programming style, a well-liked technique for handling sizable data collections. Data transport speed between nodes is one of HDFS's main characteristics. This makes it possible to process big data sets across a cluster of computers effectively [21]. The capability of HDFS to eliminate duplicate data from the collection is another crucial feature. As a result, less storage space is required and data processing efficiency is increased.HDFS does not alter the file after it has been written; instead, it retrieves data based on the file name. This guarantees the consistency and integrity of the data saved in HDFS. Therefore, the entire file must be rewritten if any modifications are necessary. To solve a query, the HDFS uses "2" phases, such as the map function and the reduced function, which are represented as follows:

$$H\ (Dn) = [Pf, Cf] \quad (2)$$

Characteristic map () The Map () function is the first feature to be had in the Map/reduction device. The master node (MN) is where this feature is dominant. The input facts is divided up or processed into a couple of smaller sub-processes. The employee nodes, which cope with those little methods, are where those sub-methods are further dispersed. The MN is then given an acknowledgment.

$$Pf = map(H(4)(Dn)) \quad (3)$$

Map () denotes the feature that plays the mapping, and Pf is the result of the map () characteristic.

The reduce () characteristic is the subsequent feature within the Hadoop tool that is important. This function gathers the thorough sub-operation consequences, combines them, after which offers the aggregated choice-based consequences as an acknowledgement of the initial primary desires. the subsequent mathematical equation shows a way to denote the discount feature:

$$Cf = (5)reduce(Pf) \quad (4)$$

Where Pf is the result of the mapping function, reduce () reduces the components, and Cf is the reduced set of data.

C. Feature extraction

Important properties together with closed frequent object set, aid, and self-belief are retrieved from the unique data after repetitive information has been removed. Ultimately, the fee of aid and self-assurance is controlled based on the entropy calculation. Using the normalized k-way (NKM) method, big data format this component used a Horse Herd technique to accomplish a BD layout. Here, it begins via the use of the guide and self-belief features' entropy fee, which incorporates of lowest and most values [22].

D. Normalization

The normalization is stated as,

$$N \propto= \frac{D - Dmin}{Dmax - Dmin} \quad (5)$$

Where Dmax and Dmin are the maximum and minimum values of the number of data D, respectively, and N stands for the normalized value of Dmax. Next, use Horse Herd to calculate the total number of clusters and beginning centroids. The Horse Herd clustering algorithm is used to divide N values into c clusters. The clustering problem is solved using the algorithm in order to determine the objective function's ideal value. Better clustering outcomes are indicated by a smaller value for the objective function [23].

The Horse Herd Optimization Algorithm (HOA) mimics the behaviour of different age groups of horse herds. Horse behaviour can be categorized into six broad divisions based on common patterns: grazing, hierarchy, sociability, imitation, defence mechanism, and roaming [22]. At

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
DOI: https://doi.org/10.17762/ijritcc.v11i10.8517
Article Received: *02 August 2023* ***Revised:*** *29 September 2023* ***Accepted:*** *13 October 2023*

different ages, horse's exhibit varied behaviours. To determine the age of the horses, a thorough matrix of responses should be constructed for each iteration.

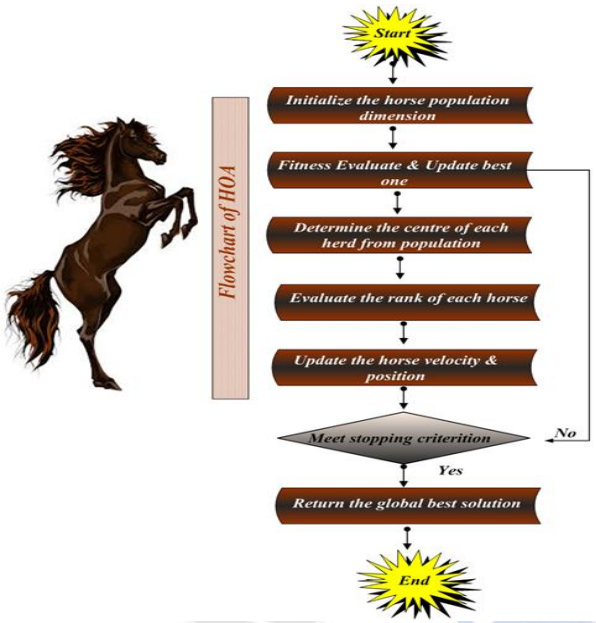

Figure 3. Flow diagram of Horse Herd Optimization Algorithm

IV. RESULTS AND DISCUSSION

This section tested the effectiveness of the counselled question optimization approach whilst applied to the pony Herd method. The device overall performance is tested and compared to the cutting-edge facts size-targeted processes. The information sizes between 10 and 50 MB are considered right here. The JAVA running platform makes use of the proposed question optimization era. Here, performance analysis is used to examine the performances of the proposed Horse Herd approach with those of the bushy C manner (FCM) and Horse Herd algorithms that are already in use.

Tables 1, 2, 3, and 4 exhibit the performance comparison among the proposed Horse Herd and the modern procedures. Discussion desk 1 in comparison the sensitivity, accuracy, and specificity-cantered performances of the proposed Horse Herd with those of the modern-day Horse Herd and FCM. Based totally on records volumes ranging from 10 to 50 MB, overall performance varies. The proposed Horse Herd clustering supplies accuracy in the range of ninety one to ninety six for all the studied information sizes. The proposed Horse Herd outperformed the prevailing approaches in all of the report sizes that were taken into consideration. It follows that the proposed Horse Herd plays at an excessive stage.

Discussion table 2 evaluated the do not forget, precision, and F-cost performance outcomes of the proposed Horse Herd with the ones of the cutting-edge FCM and Horse Herd. The proposed Horse Herd has 92.forty five% precision, zero.21% do not forget, and 89.72% F-measure for 30 MB of facts. Similar to this, the proposed Horse Herd performs better with ultimate statistics sizes like 10, 20, 30, and 50 MB. Don't forget and precision measures are used to create the F-degree metric. A system is considered to be a very good system if its F-degree is high. As a result, the cautioned Horse Herd is regarded as a good system because it offers a advanced F-measure than the current k-means and FCM.

TABLE 1.COMPARE THE ACCURACY, SPECIFICITY, AND SENSITIVITY FOR PROPOSED ALGORITHM WITH EXISTING ALGORITHM

Technique/Data size in MB		**10**	**20**	**30**	**40**	**50**
FCM [24]	***Accuracy %***	71	73	79	83	88
	Specificity %	72	77	80	82	85
	Sensitivity %	82	83	85	87	90
K-Means [25]	***Accuracy %***	77	78	82	85	89
	Specificity %	76	79	83	84	87
	Sensitivity %	86	89	90	91	92
Proposed algorithm	***Accuracy %***	80	84	87	92	95
	Specificity %	82	86	88	91	93
	Sensitivity %	92	91	94	94	97

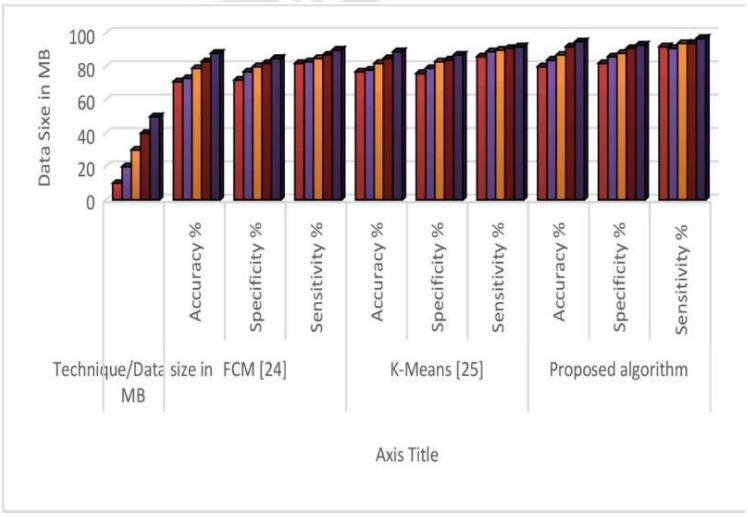

Figure 4. Compare the accuracy, specificity, and sensitivity for proposed algorithm with existing algorithm

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
DOI: https://doi.org/10.17762/ijritcc.v11i10.8517
Article Received: *02 August 2023* ***Revised:*** *29 September 2023* ***Accepted:*** *13 October 2023*

TABLE 2.COMPARISON TABLE BASED ON PRECISION, RECALL, AND F-MEASURE FOR PROPOSED ALGORITHM WITH EXISTING ALGORITHM

echnique/Data size in MB		10	20	30	40	50
FCM [24]	*Precision %*	81	83	79	88	89
	Recall %	72	77	80	82	85
	F-value %	82	83	85	87	90
K-Means [25]	*Precision %*	87	88	82	85	89
	Recall %	81	84	86	95	91
	F-value %	86	87	89	90	91
Proposed algorithm	*Precision %*	85	87	88	92	92
	Recall %	85	87	89	91	93
	F-value %	89	90	93	95	96

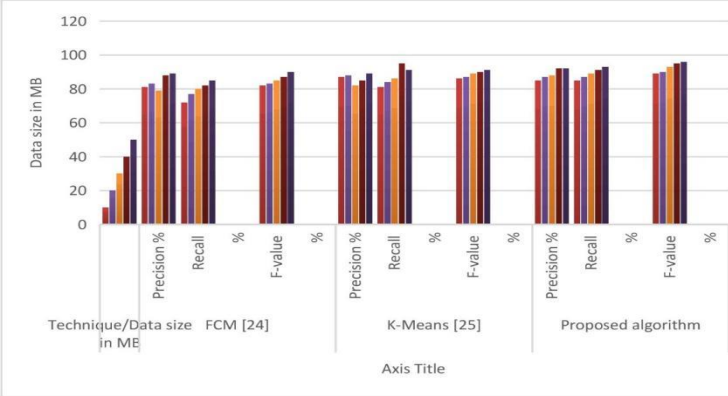

Figure 5. Comparison table based on precision, recall, and F-measure for proposed algorithm with existing algorithm

TABLE 3.TERMS OF RETRIEVAL TIME AND EXECUTION TIME FOR PROPOSED ALGORITHM WITH EXISTING ALGORITHM

Technique/Data size in MB		10	20	30	40	50
FCM [24]	*Retrieval time*	783	2155	4755	6442	8766
	Execution time	158	189	257	297	339
K-Means [25]	*Retrieval time*	443	1405	2092	4171	6677
	Execution time	122	153	224	234	305
Proposed algorithm	*Retrieval time*	255	774	1395	2333	2699
	Execution time	110	131	177	211	271

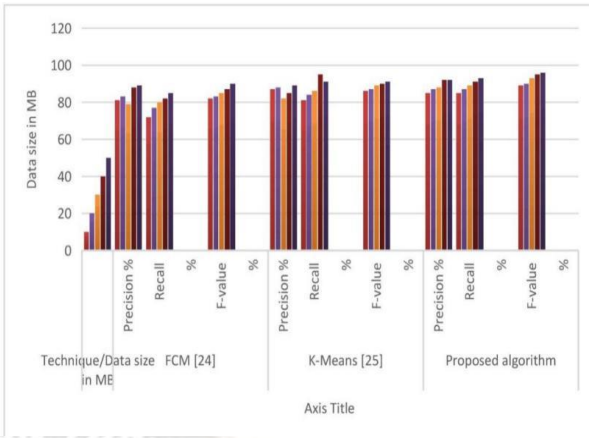

Figure 6. Terms of retrieval time and execution time for proposed algorithm with existing algorithm

TABLE 4. MEMORY PRACTICE INVESTIGATION

Technique/ Data size in MB	10	20	30	40	50
FCM [24]	42,22,465	46,66,522	50,24,000	53,66,756	55,78,443
K-Means [25]	47,77,543	50,13,345	54,66,232	58,77,445	60,34,557
Proposed algorithm	52,33,344	56,44,323	60,21,235	62,33,565	64,33,121

Discussion Table 4 contrasted the memory use based performance claims made by the proposed Horse Herd, the existing FCM, and K-Means. The suggested horse Herd requires 5366,756 kilobytes of memory storage to operate on 40 MB of data. However, the current K-Means and FCM take up 6034,557 and 6433,121 kilobytes, respectively. The proposed horse Herd uses less memory than the current approaches for the other remaining data size. As a result, it can be concluded that the suggested horse Herd achieves high-level performance compared to the existing techniques.

V. CONCLUSION

Due to the recognition of massive-scale records evaluation structures like the Hadoop machine, question optimization in business analytics (BD) is now a promising study vicinity. The usage of the HDFS map reduce method and the horse Herd set of rules, this research supplied a greater query optimization process in BD. The BD association segment and the query optimization phase are the 2 stages of the proposed paintings. Data length was used to examine the overall performance of the suggested device. The

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
DOI: https://doi.org/10.17762/ijritcc.v11i10.8517
***Article Received:** 02 August 2023 **Revised:** 29 September 2023 **Accepted:** 13 October 2023*

documents are between 10 and 50 MB in size. The contrast outcomes verified that the recommended work offers more accuracy and requires less time for query retrieval. Additionally, the suggested approach makes use of less memory area. As a result, our recommended device is superior to the modern-day device. The effectiveness of this machine can potentially be multiplied in the future through incorporating characteristic selection to hurry up retrieval and by using utilising optimization techniques.

REFERENCES:

[1]. Rawat, J.S., Kishor, S., Kumari, M.: A survey on query optimization in cloud computing. Int J AdvTechnolEngSci 4(10), 2348 (2016)

[2]. Gu, R., Yang, X., Yan, J., Sun, Y., Wang, B., Yuan, C., Huang, Y.: SHadoop: improving mapreduce performance by optimizing job execution mechanism in hadoop clusters. J Parallel DistribComput. 74(3), 2166–2179 (2014)

[3]. J Wolf, D Rajan, K Hildrum, R Khandekar, V Kumar, S Parekh, and KL Wu 2010, "Flex: A slot allocation scheduling optimizer for mapreduce workloads", In Proceedings of the ACM/IFIP/USENIX 11th International Conference on Middleware, SpringerVerlag, pp. 1-20

[4]. Barba-González, C., García-Nieto, J., Nebro, A.J., Cordero, J.A., Durillo, J.J., Navas-Delgado, I., Aldana-Montes, J.F.: jMetalSP: a framework for dynamic multi-objective big data optimization. Applied Soft Computing 69, 737–748 (2018)

[5]. Song, J., Ma, Z., Thomas, R., Ge, Yu.: Energy efficiency optimization in big data processing platform by improving resources utilization. Sustainable Computing: Informatics and Systems 21, 80–89 (2019)

[6]. Mahajan, D., Blakeney, C., Zong, Z.: Improving the energy efciency of relational and NoSQL databases via query optimizations. Sustainable Computing: Informatics and Systems 22, 120–133 (2019)

[7]. Rini John, and Nikita Palaskar, "A survey of various query optimization techniques", International Journal of Computer Applications, vol. 975, pp. 8887

[8]. Roy, C., Pandey, M., Rautaray, S.S.: A proposal for optimization of data node by horizontal scaling of name node using big data tools. In: Proceedings of the 3rd International Conference for Convergence in Technology (I2CT), IEEE, pp. 1–6 (2018)

[9]. Dwivedi, J., Tiwary, A.: Big data analytics: an overview. Int. J. Sci. Technol. Res. 5(07) (2016)

[10]. ElhamAzhir ,MehdiHosseinzadeh , Faheem Khan , and Amir Mosavi : Performance Evaluation of Query Plan Recommendation with Apache Hadoop and Apache Spark. 10(19), 3517(2022)

[11]. Deepak Kumar, Vijay Kumar Jha: An improved query optimization process in big data using ACO-GA algorithm and HDFS map reduce technique. Springer Science+Business Media, LLC, part of Springer Nature (2020)

[12]. Song, J., Ma, Z., Thomas, R., Ge, Yu.: Energy efciency optimization in big data processing platform by improving resources utilization. Sustainable Computing: Informatics and Systems 21, 80–89 (2019)

[13]. Panahi, V.; Navimipour, N.J. Join query optimization in the distributed database system using an artificial bee colony algorithm and genetic operators. Concurr. Comput. Pract. Exp. 2019, 31, e5218.

[14]. Pasquale Salza, FilomenaFerrucci. Speed up genetic algorithms in the cloud using software containers. (2019)

[15]. Mahajan, D., Blakeney, C., Zong, Z.: Improving the energy efciency of relational and NoSQL databases via query optimizations. Sustainable Computing: Informatics and Systems 22, 120–133 (2019)

[16]. Bao, C., Cao, M.: Query optimization of massive social network data based on hbase. In: Proceedings of the IEEE 4th International Conference on Big Data Analytics (ICBDA), pp. 94–97 (2019)

[17]. Sahal, R., Nihad, M., Khafagy, M.H., Omara, F.A.: iHOME: index-based join query optimization for limited big data storage. J. Grid Comput. 16(2), 345–380 (2018)

[18]. Rawat, J.S., Kishor, S., Kumari, M.: A survey on query optimization in cloud computing. Int J AdvTechnolEngSci 4(10), 2348 (2016)

[19]. KiranjitPattnaik, Bhabani Shankar Prasad Mishra: A Review on Parallel Genetic Algorithm Models for Map Reduce in Big Data. International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181 IJERTV5IS080400 Vol. 5 Issue 08, August-2016.

[20]. Panahi, V.; Navimipour, N.J. Join query optimization in the distributed database system using an artificial bee colony algorithm and genetic operators. Concurr. Comput. Pract. Exp. 2019, 31, e5218.

[21]. Rani, S.; Rama, B. MapReduce with Hadoop for Simplified Analysis of Big Data, International Journal of Advanced Research in Computer Science, May-June 2017, Volume 8, No. 5, ISSN No. 0976-5697, pp. 853-856.

[22]. Joseph, C.W.; Pushpalatha, B., A Survey on Big Data and Hadoop, International Journal of Innovative Research in Computer and Communication Engineering, ISSN(Online): 2320-9801, March 2017, Vol. 5, Issue 3, pp. 5525-5530.

[23]. Ferrucci, F., Salza, P., and Sarro, F. (2016). Using Hadoop MapReduce for Parallel Genetic Algorithms: A Comparison of the Global, Grid and Island Models - Appendix. https: //doi.org/10.6084/m9.figshare.5091898.

[24]. Fu, W., Menzies, T., and Shen, X. (2016). Tuning for Software Analytics: Is It Really Necessary? Information and Software Technology, 76:135–146.

[25]. Salza, P., Ferrucci, F., and Sarro, F. (2016a). Develop, Deploy and Execute Parallel Genetic Algorithms in the Cloud. In Genetic and Evolutionary Computation Conference (GECCO), pages 121–122.