iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** IX **Month of publication:** September 2020

DOI: https://doi.org/10.22214/ijraset.2020.31679

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue IX Sep 2020- Available at www.ijraset.com

An Efficient Clustering Algorithm with Enhanced MapReduce Design based Modified K Means for Outlier Detection

Arjun Singh Thakur[1], Pooja Trivedi[2]
[1, 2] *Computer Science and Engineering Department, RKDF School Of Engineering Indore (MP)*

Abstract: ***With an increase in amount of data also the speed of access to information has increased. Large data is the distorted form of every data well gained from different sources such as photos, , videos, social media sharing, network blogs, log files, etc. into a consequential and feasible forms. Clustering methods are very useful. Clustering process permits very analogous data to be positioned under a group by unscrambling the data into a explicit group. Once datasets are separated, outlier detection is used to discover false data. In this research work, goal is to formulate data clustering and outlier detection process quicker by using MapReduce technology with modified K-means clustering method. Clustering on enormous data can be time overriding. Hence, MapReduce computing design is utilized and focused on reliable, unfailing and swift clustering process by this technology. The successful execution with comparison to traditional K means is drawn. The results are offered in tables and graphs using sample dataset. The obtained results prove that an improved execution time is achieved in k means MapReduce algorithm along with a robust and efficient system for removing the outliers. The results also revealed an efficacy in terms of resource optimization.***
Keywords: Clustering, K-means, mapReduce, Hadoop, Outlier detection

I. INTRODUCTION

In recent times clustering is the most admired unsupervised-learning algorithm used broadly in various aspects of lives [1]. In clustering the classification of an object into a group is formed through the searching of similarities between the objects [2]. There are three main types of clustering algorithms, k means, k medoids, CLARANS. From all K means is most popular clustering algorithm and is classified under unsupervised learning algorithm [3].Besides, K-means algorithm is ranked as a top 10 algorithm in data mining [4]. Numerous of algorithm has build whose rationale is to optimize algorithm performance, for example, K-means++ and Pillar K-means. Only these algorithms approaches object initialization as cluster centroid to improve the k means algorithm performance. The parallel computing algorithms found as an answer to time complexities but expensive for big dataset with millions of objects [5]. Hence it is essentially important to develop a data-driven clustering algorithm. Therefore, in cluster computers map reduce framework can be used potentially for the processing of big data [6,7]. Map reduce is a computational standard which is implemented by Hadoop. In hadoop applications are partitioned into lots of small fragments of work, each one is executed or re executed on several node in the cluster.

1) *Map Reduce:* The map reduce framework arrange output data by key after process. Then based on similarity of key the data is entered into same reduce task which combine them to produce a single result. Map reduce is precisely explained in [8].The key benefits of Map reduce its ease of use as it is not parallel programming oriented. It is scalable through addition of servers the processing power can also be increased. Intra machine communication and machine failure are also handled by Map Reduce.
2) *Haadoop:* There are two components of hadoop

a) The Hadoop Distributed File System (HDFS)
b) MapReduce Engine

The (HDFS) Hadoop Distributed File System is an expanded comprehensive version of google file system. It is intended to accumulate reliably a huge amount of data on cluster of machines. The data is streamed at high bandwidth to user applications. Besides this map reduce is consist of a job tracker, it exists in a master node. Slave node of job tracker also has task trackers.

3) *Kmeans Clustering:* K means clustering is an unsupervised learning algorithm of data mining which is used to categorize semi structured or unstructured data sets. Due to its simplicity and efficacy to classify the voluminous datasets it is most widely used. It works on number of cluster, a parameter of intial set of centroids. With respective to the centroid of each cluster distance of each item is calculated. The recalculation of centroid of the cluster to which the item was assigned takes place. The distance can be calculated through many techniques in existence. But usually Euclidean distance is preferred. It has been observed confirmed that for real world large datasets, MapReduce is more efficient than MPI and OpenMP. The majority time taking part of k-means is the iterative distance computation. [9] found that the main focal point for gaining efficiency of algorithm is the optimization of iterative part of algorithm. Due to the necessities of using MapReduce with Hadoop distributed construction into modifying k-means attracted our focus to work on it.

II. LITERATURE SURVEY

As compared with other data mining techniques, clustering can be inclusive for the classification of data without prior knowledge. Hence this paper gives the brief summary of current literature related to data mining, map reduce and K means clustering algorithm. Nowadays, numerous data mining methods are available to identify the types of patterns to be found in data mining task. These methods comprise frequent patterns mining, discrimination and characterizations, classification and regression, correlations and associations, clustering analysis, outlier analysis [9,10]. Clustering is one of the foremost research fields in the ample area of data mining and analysis. The concept of clustering is to partition the data objects of a dataset into a number of groups or subsets such that objects in a particular subset are analogous to each other and relatively dissimilar from objects from other subsets [11]. Every subset is a single cluster formed on the notion of minimizing interclass and maximizing the intraclass similarity. On the basis of feature values that unfolds the objects and a range of distance measures this similarity and dissimilarity are evaluated. However it becomes very tough because of the constant grow of data quantity. Numerous algorithms have been designed to execute clustering, everyone comprise of different principle.

A L Ramdani and H B Firmansyah et al [5] implemented Pilar K means algorithm in a distributed system using map reduce framework. Authors explored the existing Pillar K-Means Algorithm by using MapReduce Framework with a variety of functions like Mapper and Reducer, which are component of MapReduce Framework and implemented on Hadoop. The obtained result describes that there was an optimistic performance concerning efficiency and scalability of Pillar K-means through synthetic datasets.The map reduce enables computational speed by adding number of nodes.

Savvas et al [12] The objective of the proposed work is to analyses the effectiveness of MapReduce to cluster. Performance of the developed technique is optimized with 3 different numbers of clusters with 21 nodes. The author modified the map reduce method to k means algorithm and discussed the theoretical complexities. By keeping hadoop criteria in mind the clustering response time was improved. Although the consequence of number of nodes on the speed ofconvergence of clustering technique also studied. The performance improved through lessening the quantity of in-between read/write operations and adding up a combiner among map and reduce jobs. Additionally the significant raise in performance of system was achieved by available nodes for the map tasks.

T.H. Sardar et al [13] developed and tested a parallel k-means algorithm using MapReduce programming model and compared it with sequential k-means for clustering varying size of document dataset.Authors compared the sequential k means algorithm with the clustering job execution time among the datasets of different sizes. The proposed framework discovered that it is competent to cluster datasets in short span of time by utilizing hadoop clusters of 10 nodes.

Boukhdhir et al [14] proposed an enhanced design of k-means based on mapReduce in order to adapt it to tackle large-scale datasets by sinking its execution time. In addition two more algorithms are designed to eliminate outliers from the dataset and for an automated selection of initials centroids thus alleviate the result. the proposed algorithm found more efficient, more adapted to handle large-scale datasets than existing algorithms. Additionally two limitations are present first the value of k is required as input and other it can be functional only for datasets which have attributes with numerical values.

Sreedhar et al. [15] presented two methods to the clustering of large datasets using MapReduce. K-Means Hadoop MapReduce (KM-HMR) was discovered primarily that work on the MapReduce performance with standard K-means. Another approach improves the quality of clusters to construct maximum intra-cluster and minimum inter-cluster distances for large datasets. Efficient clustering achieved with the noteworthy improvements in execution times.

T.Mohana et al [16] proposed an algorithm scaled K means that achieve enhanced results while managing clusters of circularly distributed data points and somewhat overlapped clusters. The evaluation is based on 10 different samples of data and carry outon scaled kmeans algorithms via Hadoop Map Reduce. The result illustrates an understandable sign of having the similar objects in the same cluster. This research work carried out using Hadoop and Map Reduce framework that furnishes high performance in big data analysis.

III.METHODOLOGY

It has been clear from the study of various researchers that most of the recent work focused on k means with map reduce paradigm and executed on hadoop platform to reduce the execution time to cluster dataset. Hence the major goal of the proposed research is to determine clustering efficiency in terms of execution time of modified k means and compare it with classical k means on different data sizes, accompanied with outlier detection. For this purpose our methodology consists of Modified k means algorithm with mapReduce algorithm. Map reduce is a computational standard which is implemented by Hadoop. define only two functions Map and Reduce and it is liable of all the left over task. It allows division of data into small data blocks. A map function is called for each data block. The input and output data are in the form of key-value pairs (k,v) produced by both Map and reduce stages.

The map reduce framework arrange output data by key after process. Then based on similarity of key the data is entered into same reduce task which combine them to produce a single result.

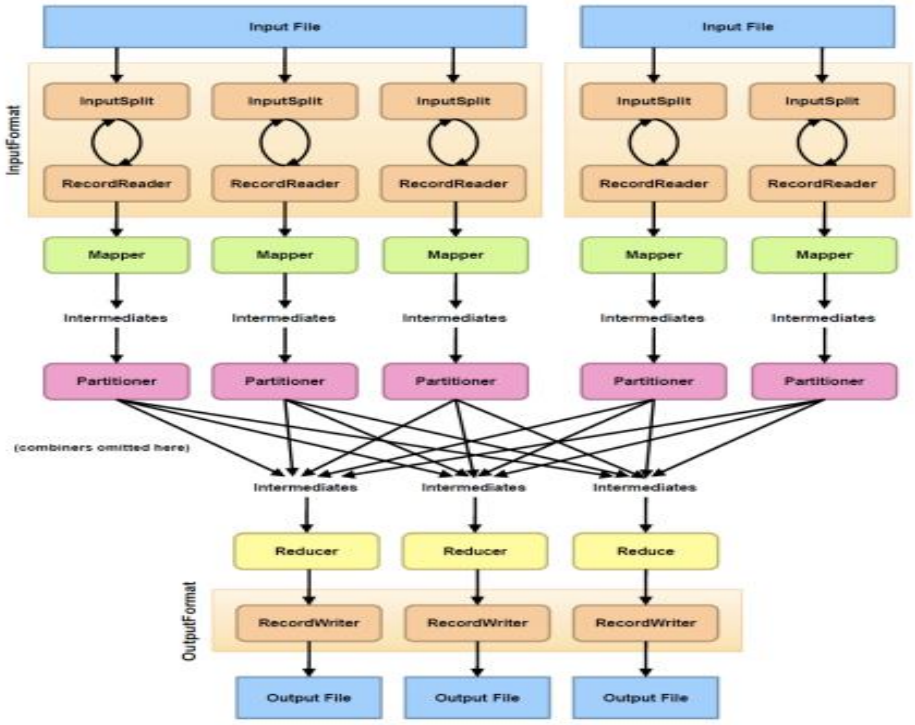


Fig.1 : Architecture of Proposed work

A. Steps in Map Reduce

Map takes a data in the form of pairs and returns a list of <key, value> pairs. This case doesn't have any unique key.

1) The Hadoop framework utilizes the results of Map, sort and shuffle simultaneously .This sort and shuffle acts on these list of <key, value> pairs and sends out unique keys and a list of values associated with this unique key <key, list(values)>.

2) The reducer phase receives the output of sort and shuffle then perform a defined function on list of values for unique keys and Final output will<key, value> will be stored/displayed.

Proposed K-Mean MapReduce algorithm divided into two sub algorithms.

a) K-Mean Mapper Algorithm

b) K-Mean Reduce Algorithm

For distance calculation between two points conventional Euclidean distance is used.

B. K-Mean Map Reduce Steps

The following steps performed in the K-Mean Map reduce clustering

1) *Step 1:* Select centroids at k random points also known as cluster centers.

a) Clusters the data into *k* groups where *k* is predefined.

b) Select *k* points at random as cluster centers.

2) *Step 2:* Apply Euclidean distance for calculation of centroids and allocate each x(i) to the nearby cluster. Where dist() is the Euclidean distance. Here, we calculate the distance of each x value from each c value, i.e. the distance between x1-c1, x1-c2, x1-c3, and so on. Then we find which is the lowest value and assign x1 to that particular centroid. In the same way, we discover the least distance for x2, x3, etc. Assign objects to their closest cluster center according to the *Euclidean distance* function.

3) *Step 3:* Calculate the centroid or mean of all objects in each cluster

4) *Step 4:* Repeat steps 2 & 3 in anticipation of convergence.

Repeat steps 2, 3 and 4 until the same points are assigned to each cluster in consecutive rounds.

C. Flowchart of Proposed Work

The flowchart below shows how Modified k-means clustering work.

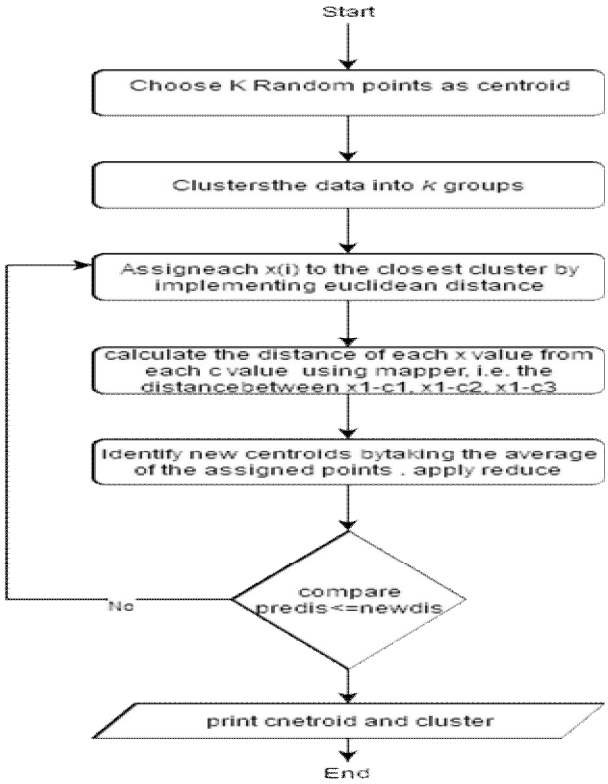


Fig 2: Flowchart of Proposed work

IV. RESULTS

Java is used for the implementation with WHO dataset. The java has been using Netbeans8.0 which provides easy to implement graphical user interface for the proposed system. Our proposed method is implemented with different dataset lengths starting from 10 objects to10,000 objects. The total execution time is recorded for different executions on Kmeans and mapReduce k means. With the obtained results it can be stated that, there is a significant improvement achieved in time of execution on different lengths of data objects. Additionally required Memory for various executions has been recorded for different steps of the proposed work and results have been drawn. The proposed method results are better with respect to K means in terms of memory requirement. The obtained result for time complexities is shown below.

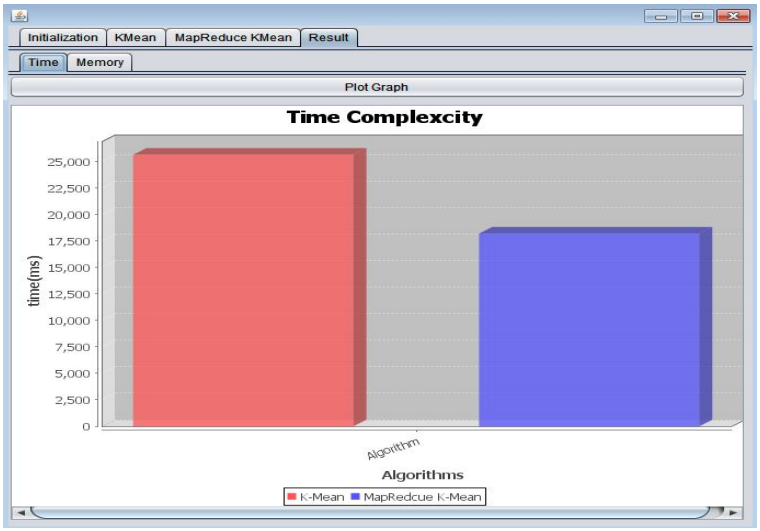


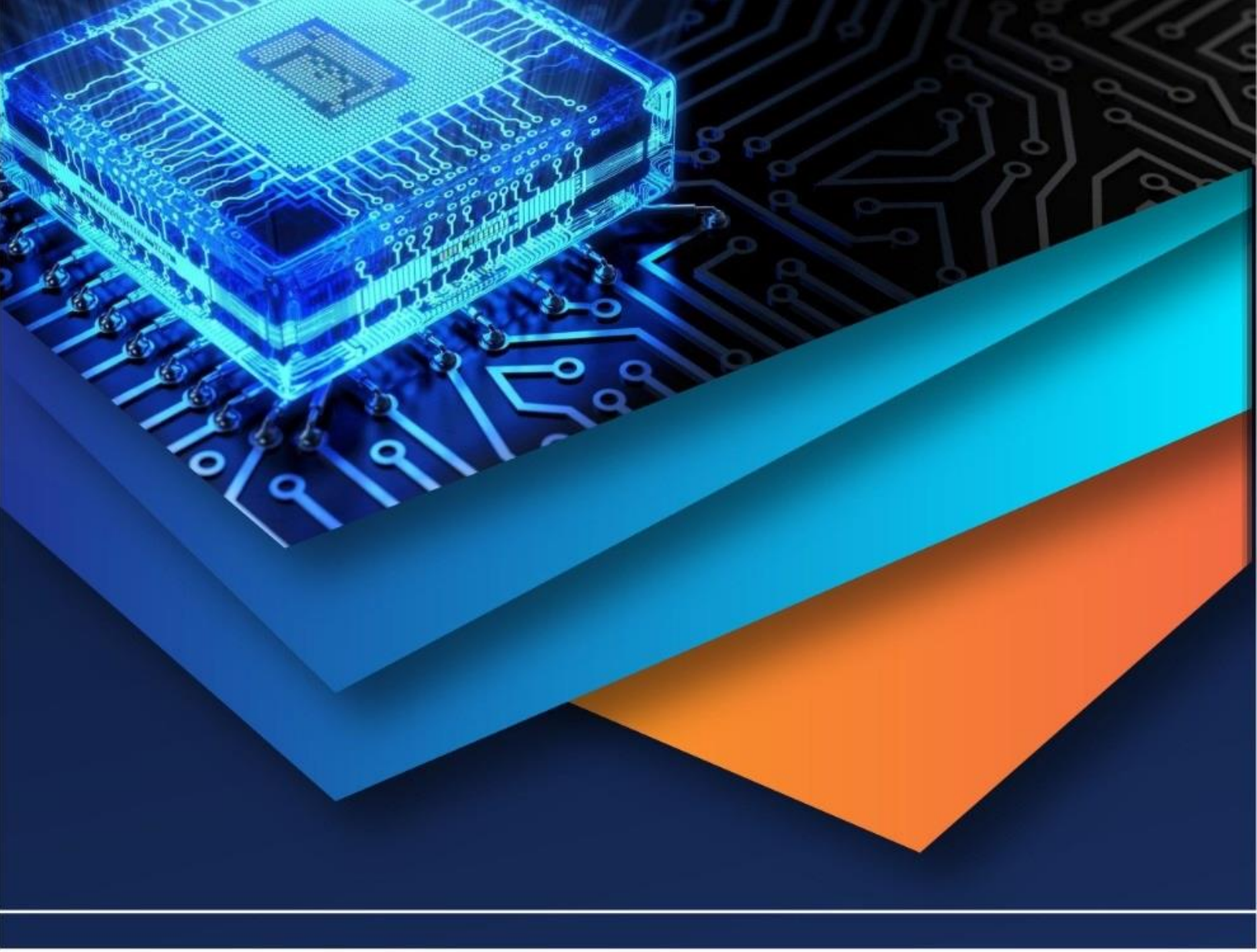
Fig 3: Comparison graph of time complexities Of K mean and Map Reduce K mean

V. CONCLUSION

In today's world, huge quantity of data processing is engaged due to voluminous exchange of information. To act in response to this necessitate we attempt and implement the fundamental algorithms used for data mining on a distributed or a parallel environment to condense the operational resources and increase the speed of operation. Implemented software has been run using various lengths of data start from a dataset of ten objects to ten thousand objects. The execution time is tested and compared with the traditional k means algorithm Memory required for various executions has been recorded for different steps of the proposed work and results have been drawn.With the help of this proposed method efficient implementation of modified K means map reduce clustering algorithm for outlier detection is attained. The obtained results prove that an improved execution time is achieved in k means MapReduce algorithm along with a robust and efficient system for removing the outliers. The results also revealed an efficacy in terms of resource optimization.

REFERENCES

[1] J. Han, M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, 2000
[2] V. Estivill-Castro, "Why So Many Clustering Algorithms-A Position Paper", SIGKDD Explorations, 2002, Vol. 4, No. 1, pp. 65-75.
[3] J. B. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations",Berkeley: University of California Press, 1967, pp. 281-297.
[4] X. Wu, V. Kumar, J.-R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. Mclachlan, A. Ng, B. Liu, P.Yu, Z.-H. Zhou, M. Steinbach, D. Hand, and D.Steinberg, "Top 10 Algorithms in Data Mining", Knowledge and Information Systems, 2008, Vol. 14, No. 1, pp. 1-37
[5] A L Ramdani and H B Firmansyah , Pillar K-Means Clustering Algorithm Using MapReduce Framework 2019 IOP Conf. Ser.: Earth Environ. Sci. 258 012031, IOP Conf. Series: Earth and Environmental Science (ICoSITeR) 2018, 258 (2019) 012031 doi:10.1088/1755-1315/258/1/012031
[6] Dean, J., Ghemawat, S.: MapReduce: Simplified Data Processing on Large Clus- ters. Communications of The ACM 51(1), 107–113 (2008)
[7] J. Dean and S. Ghemawat. Mapreduce: simplified data processing on large clusters. Proceedings of the 6th conference on Symposium on Opearting Systems Design & Implementation, 2004.
[8] T. White. Hadoop: The Definitive Guide. 1st ed. O'Reilly Media, Inc.,2009.
[9] Jinlan Tian, Lin Zhu, Suqin Zhang, Lu LIU, et al. Improvement and parallelism of k-means clustering algorithm. Tsinghua Sci Technol 2005;10(13):277.
[10] Ping ZHOU, Jingsheng LEI, Wenjun YE. Large-scale data sets clustering based on MapReduce and Hadoop. J Comput Inf Syst 2011;7(16):5956.
[11] Marisiddanagouda M, Mr Raghu MT. Survey on performance of Hadoop MapReduce optimization methods. Int J Rec Res Math Comput Sci Inf Technol 2015;2(1):114.
[12] Ilias K. Savvas and M-Tahar Kechadi , "MINING ON THE CLOUD K-means with MapReduce, International Conference on Cloud Computing and Services Science CLOSER 2012 http://closer.scitevents.org/?y=2012
[13] T.H. Sardar, Z. Ansari," An analysis of MapReduce efficiency in document clustering using parallel K-means algorithm" Future Computing and Informatics Journal 3 (2018) 200-209
[14] Amira Boukhdhir, Oussama Lachiheb, Mohamed Salah Gouider," An improved MapReduce Design of Kmeans for clustering very large datasets" 2015 IEEE ACS 12th International Conference of Computer Systems and Applications (AICCSA)
[15] Chowdam Sreedhar, Nagulapally Kasiviswanath, Pakanti Chenna Reddy, Clustering large datasets using K means modified inter and intra clustering (KMI2C) in Hadoop,journalof big data, Sreedhar et al. J Big Data (2017) DOI 10.1186/s40537-017-0087-2
[16] T.Mohana Priya, Dr.A.Saradha, An Improved K-means Cluster algorithm using Map Reduce Techniques to mining of inter and intra cluster datain Big Data analytics, International Journal of Pure and Applied Mathematics ,Volume 119 No. 7 2018, 679-690 ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version)



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY