International Journal on Emerging Technologies **10**(2): 226-234(2019)

ISSN No. (Print) : 0975-8364
ISSN No. (Online) : 2249-3255

A Study of Clustering Taxonomy for Big Data Mining with Optimized Clustering MapReduce Model

***Kamlesh Kumar Pandey*[1] *and Diwakar Shukla*[2]**
[1]*Research Scholar, Department of Computer Science and Applications,*
Dr. Hari Singh Gour Vishwavidyalaya, Sagar, (Madhya Pradesh), India
[2]*Professor & HOD,Department of Computer Science and Applications,*
Dr. Hari Singh Gour Vishwavidyalaya, Sagar, (Madhya Pradesh), India

(Corresponding author: Kamlesh Kumar Pandey)

(Received 06 May 2019, Revised 19 July 2019 Accepted 26 July 2019)
(Published by Research Trend, Website: www.researchtrend.net)

ABSTRACT: In a big data perspective, a huge dataset was generated in real time with heterogeneous nature. Big data mining is a significant idea for extracting knowledge and hidden patterns from the volume, variety, and a velocity dataset with the help of classification, clustering, and association mining. Clustering is a significant approach to data mining. All clustering taxonomy algorithms have various challenges due to the volume, variety, and velocity of big data. Distributed and parallel execution is helpful for handling scalability and performance requirement of big data mining with the help of the MapReduce programming model. This paper presents the introduction of big data, big data mining, big data storage, and traditional clustering taxonomy. From a theoretical, practical and the existing research perspective, the paper focuses on volume, variety, and velocity criteria for identifying of clustering algorithms for big data mining and designs an optimized clustering MapReduce framework for all existing cluster algorithms and this clustering framework is explained by using the K-means algorithm. This optimized clustering MapReduce framework has more scalability and accuracy as compared to the traditional clustering algorithm.

Keywords: Big Data, Big Data Mining, Clustering, Clustering Taxonomy, MapReduce, K-means.

I. INTRODUCTION

Today era, the digital technology world produces a different type of data, such as structured, unstructured and semi-structured, take from different sources like internet of things, smartphones, logs, social network, sensor machine, internet, cloud computing and other technology, by using this technology growing day-by-day and huge amount of data generated in every second and hour. Uses of all these technologies change the nature of data from simple data to big data because of these technologies based on real-time applications and heterogeneous data. Some researchers describe the speed of data generated at the present time. According to Gantz (2012) [1], and IDC report predicts says 40 Zettabyte data will be generated, consumed and simulated per day in 2020 all over the world [2], and Dobra (2014) report says 205 Exabyte's data are generated in every day in all over the world, where 95% data are organized in the form of unstructured format. Sivarajah et al. (2016) described Facebook 500 Terabytes and Walmart 2.5 Petabytes data are generated, consumed and managed in every hour all over the world [3]. A Twitter user is given 7800 tweets per second and 700 million active users are using Instagram and multimedia related applications. Facebook and Twitter are the most popular applications for Social Media applications. The monthly user of Facebook has 2 billion and Twitter has 328 million [2].

The objective of this paper is identifying a clustering algorithm for big data mining and presents an optimized clustering framework for processing on big data with high scalability and performance. In this study, the first section defines what is big data, big data characteristics, big data mining, big data storages, big data processing tools. The second section of this paper presents the traditional classification of clustering algorithms with respect to big data mining. The third section gives to summarization and comparative analysis of clustering algorithms on the base of three dimensions of big data and tries to find out which clustering algorithm is suitable for big data mining using some criteria for volume, variety, and velocity. In the fourth section, this paper presents optimized MapReduce clustering framework for big data mining and explain their working process with the help of K-means algorithm and this section clearly defines the MapReduce based clustering algorithm is more suitable to sequentially clustering algorithm under big data mining.

A. Big Data

A common difference between traditional data and big data, big data growing rapidly with including the heterogeneous type of data, such as structured, unstructured and semi-structured format with complex nature but traditional data have a small volume with the specific data format. Doung Laney (2001) suggested three dimensions of big data which is known as main characteristics 3V's of big data. First V's is volume, which refers to the size of data. Second V's is variety, which refers to the heterogeneous sources of data as well as the data type such as structured, unstructured and semi-structured data.

The third V's is velocity, which refers to the speed of data generation, mining, and analysis of real-time or non-real-time data set. The first paragraph of this paper shows some survey, which is identified in the volume, variety, velocity. Social Media is the best example of 3V's of big data because social media is generating high volume data in the form of high variety in the form of high velocity. Fourth V's is Veracity introduced by IBM, Fifth V's is Value introduced by Oracle and Six V's is Variability introduced by SAS (Statistical Analysis System). Variability defined complication of datasets and defines data definition, behaviors, and structures whose meanings are continuing changes to per second, minutes and hours. The value defined particular long-range values from attributes of big data for data mining or data analysis. Veracity represents the accuracy, certainty, precision, relevance, predictive value, and quality of mine and analytical data. Seven V's is Visualization represents the mined data according to end-users as stable and readable nature [1-6].

Conceptually, the first three V's of big data are related to actually work for big data, such as data collection, storage, and transmission because big data frequently build-up from the high volume of different type of data, which is gathered from multiple sources at high speed, and last four V's supports to big data analysis for decision making [7,8]. Fig. 1. Show all the characteristics of big data [9].

Some researchers, communities, and organizations define big data. McKinsey global (2011) described Big Data as "Data whose scale, distribution, diversity, and/or timeliness require the use of new technical architectures and analytics to enable insights that unlock the new sources of business value" [10].

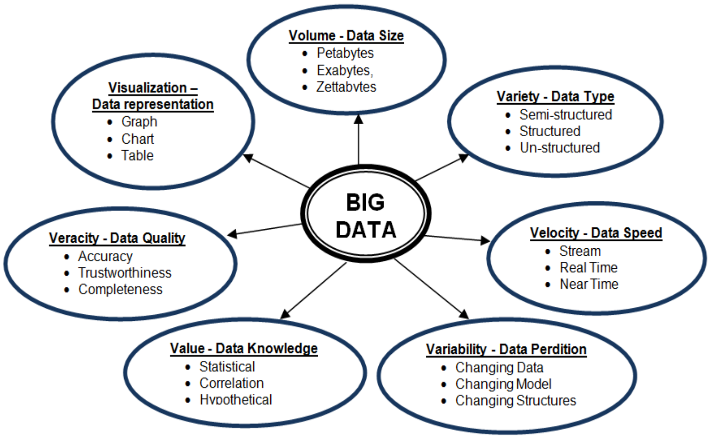

Fig. 1. 7 V's of Big Data.

National Institute of Standards and Technology defines big data as "Big data shall mean the data of which the data volume, acquisition speed, or data representation limits the capacity of using traditional relational methods to conduct effective analysis or the data which may be effectively processed with important horizontal zoom technologies" [11]. U.S. Congress (August 2013) defines big data as "large volumes of high-velocity, complex and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management, and analysis of the information" [8].

B. Big Data Mining

Nowadays huge amounts of data are collected from different fields and source by organizations and government for future used. The purpose of dataset collection is taken to decision in particular problem through data analysis, and mining. In big data mining perceptive, organizations and government are facing the biggest challenges respect to huge memory for data storage, high speed of processing and effective mining techniques [27]. Big data mining is dealt with structured, unstructured and semi-structured data, and capable of executing any query in parallel and distributed the database. Data, process, and management related challenges of big data mining motivate to develop a new effective mine algorithm [12].

Big data mining is the technique of discovering interesting hidden patterns, relations, and knowledge from high-volume, high-variety, and high-velocity dataset in parallel and distributed execution for streaming data. Big Data mining is a combination of statistics and machine learning with a streaming database management approach [6]. Traditional data mining techniques are divided in supervised (classification), unsupervised (clustering) and association mining (frequent item sets mining) but these techniques have lack of efficiency, speed, scalability, heterogeneous and accuracy when applied to Big Data Sets in a real-time environment [3, 28, 29]. The big data mining approach usually divided into three groups such as predictive, descriptive and prescriptive analysis. Predictive analyses predict the uncover value, pattern, and capture relationships in bases of current and historical data with the help of statistical method supervised and unsupervised algorithm. The descriptive analysis is given to summarize or reporting and description of the hidden patterns or relation of heterogeneous data and provide the dependencies and rules for data with the help of statistical method, for example, cluster analysis and association. The prescriptive analysis determines the relationship between data mining result and optimization policies for decision making [1,5, 30].

C. Big Data Storage for Mining

The traditional Relational database model is not sufficient for big data because of the large volume and heterogeneous data are not supported by the Relational database model. Storage of big data is classified into three bottoms up levels as file systems, databases, and programming models [6,11,13]. In file system perspective, Google developed GFS (Google file system) for large-scale and distributed applications, Hadoop developed extend version of GFS as HDFS (Hadoop Distributed File System), Microsoft developed Cosmos File System for searching and advertisement business support, Facebook developed a Haystack file system for a large number of photo storages, and Taobao developed TFS (Taobao File System) and FastDFS (fast Distributed File System).

In the database perspective, NoSQL databases the most popular database for big data storages. NoSQL database is divided into four groups based on storage technology. First one is the Key-value based database, where data are storages in the form of the value of a corresponding unique index key. This database has high expandability and takes short query response time as compared to relational databases. Some Key-value Databases are Amazon Dynamo, LinkedIn Voldemort, Redis, Tokyo Cabinet, Tokyo Tyrant, Memcached, Riak, Scalaris and so on. Second is column-oriented Database, where data is stored and processed according to columns rather than rows. Columns and rows are fractured in multiple nodes for expandability. Some column-oriented databases are BigTable inspired by Google and used to GFS, and Cassandra (Facebook). The third is a document-oriented database, where data is stored in the form of the document. This type of database basically used to semi-structured data. Some document-oriented databases are MongoDB, SimpleDB, and CouchDB. Fourth is the Graph-based database, where data is stored in a combination of vertex and edges. This type of database used to graph theory and its associated algorithm. Some graph-based databases are Neo4j, GraphDB, InfoGrdi.

Big data are generally stored and proceed by several servers. In big data mining, perspective programming models must be supported to parallel and distributed computation with NoSQL databases. MapReduce and Dryad are most popular programming model for big data mining because it has parallel and distributed access capability. In MapReduce programming model has two functions as Map and Reduce. The map function takes input in the form of key-value pairs and generates intermediate key-value pairs and Reduce function combine all intermediate values related to the same key. The Dryad programming model is a distributed and parallel execution engine of coarse-grained data. The Dryad functioning structure is a directed acyclic graph, where vertexes denote programs and edges denote data channels. It executes on the vertex in the form of clusters and transmits to data via share approaches such as channels, TCP connections, and shared-memory.

II. TRADITIONAL CLUSTERING ALGORITHM TAXONOMY FOR BIG DATA MINING

Clustering is an unsupervised technique under machine learning and data mining. The core objective of clustering is to group similar data object and separate them dissimilar data objects based on unlabelled data [14]. Webster (Merriam-Webster Online Dictionary, 2008) describes cluster analysis as “a statistical classification technique for discovering whether the individuals of a population fall into different groups by making quantitative comparisons of multiple characteristics” [15]. Clustering is the outstanding algorithm for practical perspective in big data mining such as web mining, text mining, scientific data investigation with critical observations, multimedia database, spatial database, and other database applications, Web mining or web-based analysis, medical diagnostics, CRM and marketing, computational biology and many other fields [16].

Similarity and dissimilarity are the base property for construction of the cluster. Distance function recognizes the relationship between data for creation on the cluster. The similarity feature is created via cluster through data features and natures. Some popular distance function is Euclidean Distance, Squared Euclidean Distance, Manhattan Distance, Maximum Distance, Minkowski distance, Cosine distance, Pearson correlation distance, Mahalanobis distance and so on. Some popular similarity function is Jaccard similarity, Hamming similarity, and so on [17-19].

The Clustering algorithm usually classifies into partition, hierarchical, density, grid, model, fuzzy and graph-based clustering taxonomy based on their working process, behaviors and cluster nature. Their basic cluster creation, merits, and demerits details are given in subsection A to Gof the second section [14-20].

A. Partitioning Based Clustering

These clustering methods divided the dataset into K partition based on the distance function. This method is also known as Centroid-based clustering because each K cluster partition finds the center of the cluster and assign data into the nearest center point. In general, this clustering algorithm takes low time performance, complexity, and high computing efficiency, but it's not suitable for non-convex shape data. Some typical algorithms of this kind of clustering algorithm are k-Mean, K-Medoids, PAM (Partitioning around medoids), CLARA (Clustering Large Applications), and CLARANS (Clustering Large Applications based upon Randomized Search).

B. Hierarchical Based Clustering

These clustering methods divided the dataset into the tree by Agglomerative and Divisive method, where each node represents a cluster. This method is also known as connectivity based clustering approach because the cluster is constructed by a hierarchy of clusters. The agglomerative approach starts from the bottom to top and divisive starts from the top to bottom. In agglomerative approach, assign to separate cluster for each data and these clusters are combined through distance function. The flow of this process is done by the bottom to top. In the Divisive approach, first all the data are grouped into one cluster and after this cluster is divided into different pairs of clusters as K cluster requirements based on the distance function. The flow of this process is done by top to bottom. As a practical perspective, Hierarchical Clustering is represented by the tree structure dendrogram. In general, this algorithm is suitable for arbitrary shape data, detection on easily relationship between the cluster and high scalability, but it takes the high time complexity and the number of clusters needed to be preset. Some typical algorithms of this kind of clustering algorithm are BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies), CURE (Clustering Using Representatives), ROCK (RObust Clustering uses linKs), Chameleon, ECHIDNA, WARDS, and SNN.

C. Density Based Clustering

These clustering methods divided the dataset into based on the high density of the data space. This algorithm is known as the one-scan algorithm and it is found arbitrarily shaped clusters with handle noise capability. In general, this algorithm is suitable for high efficiency and arbitrary shape data, but it takes huge memory, low quality of clustering and highly sensitive to the parameter values. Some typical algorithms of this kind of clustering algorithm are DBSCAN (Density Based Spatial Clustering of Applications with Noise), OPTICS (Ordering Points To Identify the Clustering Structure), Mean-Shift, DENCLUE (DENsity based CLUstEring), and GDBSCAN.

D. Grid Based Clustering

These clustering methods divided the data spaces into rectangular by using a hierarchical structure. The core idea of the grid clustering algorithm is that the original data space is converted into a grid format which defines the size for clustering. In general, this algorithm is suitable for high scalability, parallel processing with low time complexity, but it reduces the quality and accuracy of clusters and it is sensitive to the high granularity. Some typical algorithms of this kind of clustering algorithm are STING (statistical information Grid approach), CLIQUE, Wave Cluster, OptiGrid, MAFIA, ENCLUS, PROCLUS, ORCLUS, and STIRR (Sieving through Iterated Relational Reinforcement).

E. Model Based Clustering

These clustering methods divided the data set into based on computational models such as mathematics, and statistical distribution. This clustering algorithm is used for optimization using a mathematical model such as a probability distribution. In general, this algorithm is suitable for developing a new model with the help of existing model with describing data adequately, but it takes the high time complexity and clustering result is very sensitive to the parameters for the selected model. Some typical algorithms of this kind of clustering algorithm are COBWEB, SLINK, SOM (Soft-Organizing feature Map), ART, and EM (Expectation Maximization).

F. Fuzzy Based Clustering

These clustering methods divided the dataset into based on the discrete value [0, 1]. In general, this algorithm is suitable for high accuracy with high probability, but it has low scalability, local optimal, clustering sensitive to the initial parameter and the number of clusters needed to preset. Some typical algorithms of this kind of clustering algorithm are FCM (Fuzzy C-Means), FCS and MM.

G. Graph based Clustering

These clustering methods divided the dataset into based on the related vertex. Nodes define as the data point and edges define the relationship between data points. In general, this algorithm suitable for high efficiency with high accuracy, but time complexity depended upon the graph complexity. Some typical algorithms of this kind of clustering algorithm are CLICK and MST.

III. COMPARATIVE ANALYSIS OF CLUSTERING TECHNIQUES FOR BIG DATA MINING

The Designing of clustering algorithms for the big data mining, it needs some criteria such as volume, velocity, and variety with the efficiency of the clustering [19-21].

Volume related criteria are defining the size, high dimensional and noisy of the dataset. The size of the dataset is defined to clustering algorithm is the capability for large dataset analysis, the huge dimensions defined clustering algorithm are handle complex relations between an attribute and a large number of structures and handling of noisy or outlier defined clustering algorithm is applicable to noisy data handling itself or not.

Variety related criteria define dataset categorization and clusters Shape. The dataset categorization defines clustering algorithm can be handled which type of data, such as numeric and categorical to the physical world and the cluster shape defines a clustering algorithm should be able to take on advantages of actual data and their extensive range.

Velocity related criteria define complexity, scalability, and performance of the clustering algorithm during the execution of real dataset. The time complexity defines how much time is taken by clustering algorithm and performance respect to execution time for real data and the scalability defines when dataset size is increased, then the performance of time complexity is also roughly scaled according to dataset size.

The aim of this section is to identify clusters algorithm for three-dimensional of big data. Table 1 presents clustering taxonomies with their clustering algorithms and defines which clustering algorithm is suitable for big data mining respect of dataset size, high dimensional, handling noisy data, type of the dataset, the shapes of the dataset, time complexity, and performance and scalability criteria for volume, variety, and velocity.

IV. OPTIMIZED MAPREDUCE CLUSTERING FRAMEWORK FOR BIG DATA MINING

Every clustering algorithm has own merits and demerits based on their working process. Clustering algorithms are only suitable for big data mining if they handle any two V's of big data such as volume, variety, and velocity in general. Table 1 is identifying a good clustering algorithm chosen for big data mining using some criteria. Any clustering algorithm is working under huge dataset or high dimensional with scalability and heterogeneous data in the form of an arbitrary shape suitable for big data mining [20-22]. The introduction section clearly defines which type of database and programming model used for big data mining. Designing of a clustering algorithm for big data mining has a capability for parallel and distributed computing. Hadoop is one of the tools for the implementation of big data mining using the MapReduce programming model [2,6,12].

Various researchers have presents better to better clustering taxonomy with a specific criterion such as scalability, handling noisy data, large data set, high dimensional data, and parallel computation and so on. In this section, this paper presents an optimized clustering framework model for the execution of existing clustering algorithms using MapReduce programming model. Proposed optimized MapReduce based clustering algorithm flow chart shows in fig 2 and the model algorithm is,

Step 1: Big data set is transformed into <key, value> pairs because MapReduce used to HDFS for parallel and distributed computing.

Table 1: Summarization of Clustering Algorithm based on Three-Dimensional Properties of Big Data [14-27].

Clustering Algorithm	Volume			Variety		Velocity	
	Dataset size	High dimensional data	Handling Noisy data	Dataset type	Cluster shape	Scalability	Time complexity
Partition based Clustering							
K-Means	Large	No	High	Numerical	Convex	Medium	0 (knt)
K-Medoies	Small	No	Low	Categorical	Convex	Low	$0(k(n-k)^2)$
PAM	Small	No	Low	Numerical	Convex	Low	$0\ (k^3 * n^2)$
CLARA	Large	No	Low	Numerical	Convex	High	$0(ks^2+k(n-k))$
CLARANS	Large	No	Low	Numerical	Convex	Medium	$0(n^{2)}$
Hierarchical based Clustering							
BIRCH	Large	No	Low	Numerical	Convex	High	0(n)
CURE	Large	Yes	High	Numerical	Arbitrary	High	$0(n^2 logn)$
ROKE	Small	Yes	Low	Numerical/Categorical	Arbitrary	Medium	$0(n^2 logn)$
Chameleon	Small	No	Low	All type data	Arbitrary	High	$0(n^2)$
ECHIDNA	Large	No	Low	Multivariate	Convex	High	$0(nb(1+log_b m)$
WARDS	Small	No	Low	Numerical	Arbitrary	Medium	-------------
SNN	Small	No	Low	Categorical	Arbitrary	Medium	$0(n^2)$
Density based Clustering							
DBSCAN	Large	No	Low	Numerical	Arbitrary	Medium	0(nlogn)
OPTICS	Large	No	Low	Numerical	Arbitrary	Medium	0(nlogn)
Mean-shift	Small	No	Low	Numerical	Arbitrary	Low	0 (kernel)
DENCLUE	Large	Yes	High	Numerical	Arbitrary	Medium	0(log |d|)
GDBSCAN	Large	No	Low	Numerical	Arbitrary	Medium	---------------
Grid based Clustering							
STING	Large	Yes	Small	Spatial	Arbitrary	High	0(n)
CLIQUE	Small	Yes	Medium	Numerical	Convex	High	$0(n+k^2)$
Wave Cluster	Large	No	High	Spatial	Arbitrary	Medium	0(n)
OptiGrid	Large	Yes	High	Spatial	Arbitrary	Medium	0(nd) to 0(nd-log n)
MAFIA	Large	No	High	Numerical	Arbitrary	High	$0(c^p + p^n)$
ENCLUS	Large	No	High	Numerical	Arbitrary	High	$0(nd+m^d)$
PROCLUS	Large	Yes	High	Spatial	Arbitrary	Medium	0(n)
ORCLUS	Large	Yes	High	Spatial	Arbitrary	Medium	$0(d^3)$
STIRR	Large	No	Low	Categorical	Arbitrary	Medium	0(n)
Model based Clustering							
COBWEB	Large	No	Medium	Numerical	Arbitrary	Medium	$0(n^2)$
SLINK	Large	No	Medium	Numerical	Arbitrary	Medium	$0(n^2)$
SOM	Small	Yes	Low	Multivariate	Arbitrary	Low	$0(n^2 m)$
ART	Large	No	High	Multivariate	Arbitrary	High	(type+layer)
EM	Large	Yes	Low	Spatial	Convex		0(knp)
Fuzzy based Clustering							
FCM	Small	No	High	Numerical	Convex	Medium	0(n)
FCS	Small	No	High	Numerical	Arbitrary	Low	0 (kernel)
MM	Small	No	low	Numerical	Arbitrary	Low	$0(v^2 n)$
Graph based Clustering							
CLICK	Large	No	High	Numerical/Categorical	Arbitrary	High	0(kf(v,e))
MST	Large	No	High	Numerical/Categorical	Arbitrary	High	0(e log v)

Step 2: The Mapper function takes <key, value> pairs as input and executes the existing cluster algorithm in parallel.

Step 3: Reduce function takes the output from Map function and combine all values based on the key and calculate cluster working criteria based upon existing clustering algorithm.

Step 4: Compare reduce cluster new criteria and map cluster criteria based on the key. If reduce cluster new criteria is not sufficient according to the map cluster criteria, then go to step 3 otherwise go to step 5.

Step 5: Reduce function given the accurate and unique number of clusters based upon existing clustering algorithm.

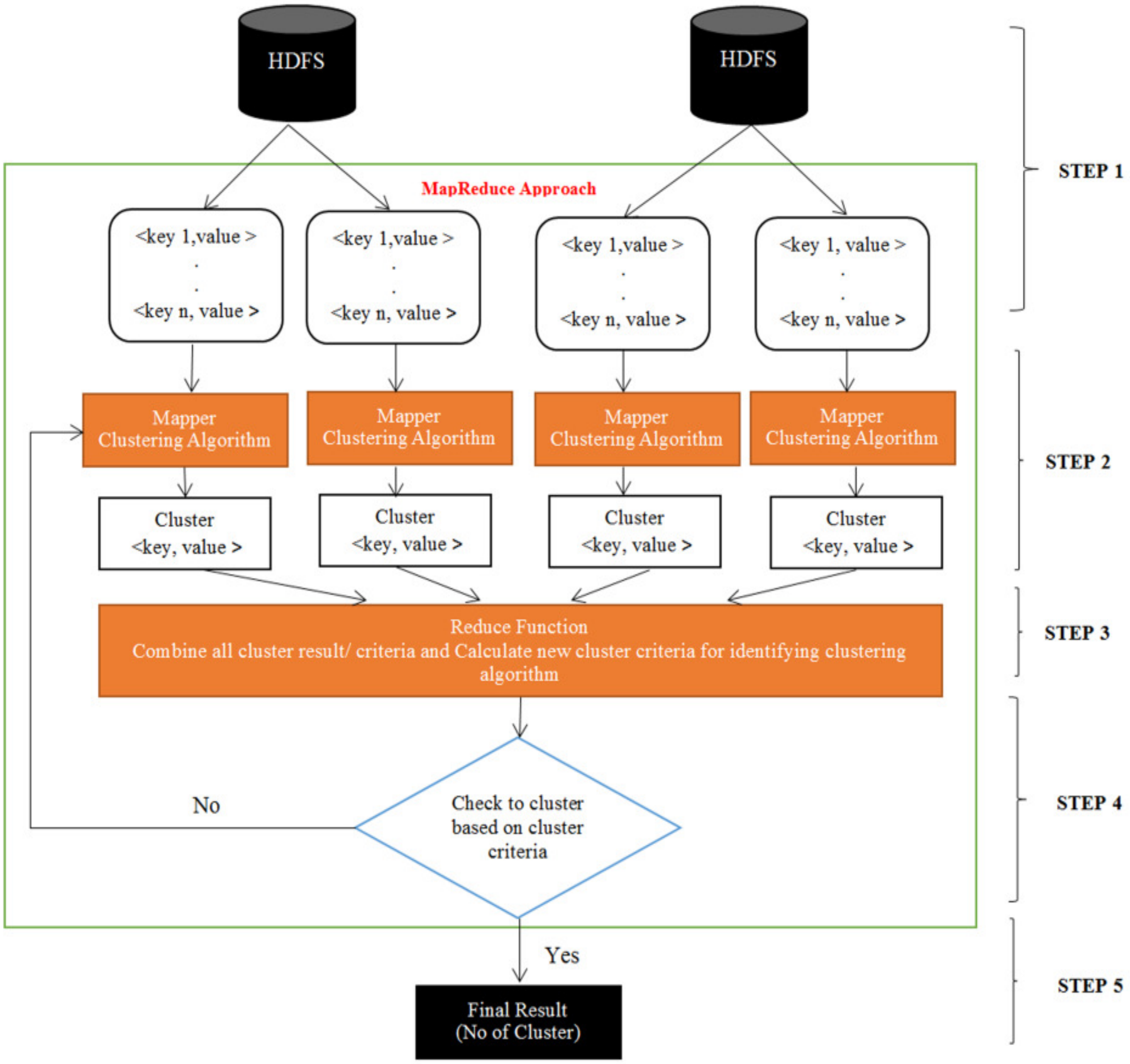

Fig. 2. Optimized Stages of clustering framework flow chart.

K-Means, BIRCH, CLARA, CURE, DBSCAN, DENCLUE, Wave cluster are good clustering algorithm for big data mining because it fulfills the criteria of big data [20]. Execute this optimized model algorithm or clustering framework with the help of K-Means clustering algorithm because of K-Means clustering capable for large dataset execution with scalability. K-Means algorithm is the top second algorithm for data mining technique. The K Means algorithm used Euclidian distance function for creating the K cluster [18, 25,26]. Execution of the K-Means algorithm is shown in table 2 using the optimized MapReduce clustering framework.

This paragraph explains the experimental interface for optimized MapReduce based clustering framework. For the experimental interface purpose, this paper used to Power dataset for creating a cluster using the K-Mean algorithm with and without this proposed optimized framework. Power dataset consists of 512,320 real data points with 7 dimensions [16]. This experiment done by using the Hadoop tool with ten nodes cluster and the system is configured as Intel I3 processor, 4 GB DDR3 RAM, 320 GB hard disk and windows 7. Table 3 shows the execution time of existing K-Means (without MapReduce) and MapReduce based K-Means clustering algorithm.

Table 2: K-Means Clustering Algorithm using Optimized MapReduce based clustering framework.

	K-Means Clustering Algorithm using Optimized MapReduce based clustering framework	
Step 1	**Input :** D Dataset (d_1, d_2 ...., d_n) K number of Cluster (k_1, k_2 ...., k_n) Distance(x, y) = Euclidian distance function pair <key, value> = D where key – index/ offset of data D value- the content of Dataset D	The big data set is transformed into <key, value> pairs.

Step 2	**Map function**
	for i=1 to key.length do
	center=null
	distmin=infinite
	For all pair.value of K do
	dist=distance(d_i , k_i)
	if (center=null or dist<dist) then
	distmin=dist
	center= k_i
	end if
	end for
	return output<center, d_i>
	end for
		The mapper function takes<key, value> pairs as input and performs the K-Means algorithm. Find out the best center based on Euclidian distance function. (The center is a criterion for cluster creation on using K-Means algorithm)

Step 3	**Reduce function**
	Input : pair <center, d_i>
	where
	center =<key>
	d_i=<value>
	centroid={}
	newcenter=null
	for i=1 to pair.length
	oldkey=pair.center
	oldvalue=pair. d_i
	centroid=oldvalue
	for end
	for i=1 to centroid.length do
	for all oldvalue of centroid do
	sumvalue=sumvalue+oldvalue
	numvalue=numvalue+1
	end for
	newcenter=sumvalue/numvalue
	returnoutput<oldkey, newcenter>
	end for
		The reduce function takes <key, value> pairs output from the map function and combine all value and sort them and find new centroid values, which is a criterion of the K-Means algorithm.

Step 4	pair <oldkey, newcenter>
	where
	oldkey=key
	newcenter=value
	for i=1 to centroid.length do
	if (newcenter= oldkey) then
	if (oldkey.key=oldvalue.value) then
	cluster well define return cluster for step 5
	end if
	Go to step 2
	end if
	end for
		First, compare both new centroid value and old centroid value for identifying the same value. If both centroid keys define the same value, then the cluster is more accurate otherwise calculated new distance through a Map function.

Step 5	Collect cluster based on pair <oldkey, newcenter>	

Table 3: Running Time of K-mean Algorithm.

Algorithm	Execution time in second
K-mean (existing)	64
K-mean (optimizedMapReduce Based)	14

V. CONCLUSION

This paper reviews the core idea of big data, big data mining, big data storage structures, execution model, clustering taxonomy and their algorithm, and defined which traditional clustering algorithm is suitable for big data mining. The first section gives information about big data, big data mining, big data, database, and comparison between traditional storage and big data mining storage technique. The second section defines the concept of all existing cluster taxonomies and their working process and behaviors such as partition, hierarchical, density, grid, model, fuzzy and graph-based clustering algorithm. The third section states the summarization of clustering algorithm based upon 3 V's of big data such as volume, variety, and velocity criteria and identifying on clustering algorithm for big data mining on the bases of size of the dataset, high dimensional, Noises in the dataset, dataset categorization, clusters Shape, complexity, and scalability. The fourth section presents a clustering framework for execution on existing clustering algorithm for using the optimized MapReduce programming model and explains their working process with the help of K-mean algorithm. This optimized clustering framework run under parallel and distributed computation, that reason it takes less time as compared to the traditional clustering algorithm.

VI. FUTURE SCOPE

This work offers the identification of clustering algorithms for big data mining and their execution through optimized MapReduce Model. The future scope of this work is executed each clustering algorithm for using this optimized MapReduce Model and validated through speedup, accuracy, precision and other clustering measures.

ACKNOWLEDGEMENT

I would like to express my very great appreciation to Prof. Diwakar Shukla, Department of Computer Science and Application, Dr. Hari Singh Gour Vishwavidyalaya, Sagar for his valuable and constructive suggestions during the planning and development of this research work. His willingness to give his time so generously has been very much appreciated.

CONFLICT OF INTEREST: Nil

REFERENCES

[1]. Sivarajah, U., Kamal, M.M., Irani, Z., & Weerakkody, V. (2017). Critical analysis of Big Data challenges and analytical methods. *Journal of Business Research,* **70**, 263-286. doi:10.1016/j.jbusres.2016.08.001

[2]. Oliverio, J. (2018). A Survey of Social Media, Big Data, Data Mining, and Analytics. *Journal of Industrial Integration and Management,* **3**(1), 1850003(1)-850003(13).doi:10.1142/S2424862218500033

[3]. Oussous, A., Benjelloun, F., Lahcen, A.A., & Belfkih, S. (2018). Big Data technologies: A survey. *Journal* of *King Saud University - Computer and Information Sciences,* **30**(4), 431-448. doi:10.1016/j.jksuci.2017.06.001

[4]. Gole, S., & Tidke, B. (2015). A survey of big data in social media using data mining techniques. In *Proceedings of IEEE International Conference on Advanced Computing and Communication Systems. IEEE Xplore Digital Library* . doi:10.1109/icaccs.2015.7324059

[5]. Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management,* **35**(2), 137-144. doi:10.1016/j.ijinfomgt.2014.10.007

[6]. Emani, C.K., Cullot, N., & Nicolle, C. (2015). Understandable Big Data: A survey. *Computer Science Review,* **17**, 70-81. doi:10.1016/j.cosrev.2015.05.002

[7]. Wasastjerna, M.C. (2018). The role of big data and digital privacy in merger review. *European Competition Journal,* **14**(2-3), 417-444. doi:10.1080/17441056.2018.1533364

[8]. Sarkar, B.K. (2017). Big data for secure healthcare system: A conceptual design. *Complex & Intelligent Systems,* **3**(2), 133-151. doi:10.1007/s40747-017-0040-1

[9]. Pandey, K.K., & Shukla, D. (2018). Mining on Relationship in Big Data era Using Apriori Algorithm. *In Proceedings of National Conference on DAMLS* (pp. 55-60.). GGU, India. ISBN: 978-93-5291-457-9

[10]. Elshawi, R., Sakr, S., Talia, D., & Trunfio, P. (2018). Big Data Systems Meet Machine Learning Challenges: Towards Big Data Science as a Service. *Big Data Research,* **14**, 1-11. doi:10.1016/j.bdr.2018.04.004

[11]. Chen, M., Mao, S., & Liu, Y. (2014). Big Data: A Survey. *Mobile Networks and Applications,* **19**(2), 171-209. doi:10.1007/s11036-013-0489-0

[12]. Bendechache, M., Tari, A., & Kechadi, M. (2018). Parallel and distributed clustering framework for big spatial data mining. *International Journal of Parallel, Emergent and Distributed Systems*, 1-19. doi:10.1080/17445760.2018.1446210

[13]. Weichen, W. (2016). Survey of Big Data Storage Technology. *Internet of Things and Cloud Computing*, **4**(3), 28-33. doi:10.11648/j.iotcc.20160403.13

[14]. Kaur, P. and Kaur, K. (2017). Comparative Study of Techniques and Issues in Data Clustering. In Saini H., Sayal, R., Rawat, S. (eds), *Innovations in Computer Science and Engineering* (Vol. **8**, Lecture Notes in Networks and Systems, pp. 1-7). Springer, Singapore.doi:10.1007/978-981-10-3818-1_1

[15]. Nagpal, A., Jatain, A., & Gaur, D. (2013). Review based on data clustering algorithms. *In Proceedings of IEEE Conference On Information And Communication Technologies (pp. 298-303). IEEE Xplore Digital Library*. doi:10.1109/cict.2013.6558109

[16]. Berkhin, P. (2006). A Survey of Clustering Data Mining Techniques. In Kogan J., Nicholas C., Teboulle M. (eds), Grouping Multidimensional Data (pp. 25-71). Berlin, Heidelberg: Springer. doi:10.1007/3-540-28349-8_2.

[17]. Chen, W., Oliverio, J., Kim, J.H., and Shen, J. (2018). The Modeling and Simulation of Data Clustering Algorithms in Data Mining with Big Data. *Journal of Industrial Integration and Management,* **12**(4): 1-16. doi:10.1142/s2424862218500173

[18]. Xu, R., & Wunschii, D. (2005). Survey of Clustering Algorithms. *IEEE Transactions on Neural Networks*, **16**(3), 645-678. doi:10.1109/tnn.2005.845141.

[19]. Xu, D., & Tian, Y. (2015). A Comprehensive Survey of Clustering Algorithms. *Annals of Data Science,* **2**(2): 165-193. doi:10.1007/s40745-015-0040-1.

[20]. Pandove, D., & Goel, S. (2015). A comprehensive study on clustering approaches for big data mining. *In Proceedings of IEEE 2nd International Conference on Electronics and Communication Systems* (pp. 1333-1338). *IEEE Xplore Digital Library*. doi:10.1109/ecs.2015.7124801

[21]. Fahad, A., Alshatri, N., Tari, Z., Alamri, A., Khalil, I., Zomaya, A.Y., Bouras, A. (2014). A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis. *IEEE Transactions on Emerging Topics in Computing*, **2**(3), 267-279. doi:10.1109/tetc.2014.2330519

[22]. Jain, A.K., Murty, M.N., & Flynn, P.J. (1999). Data clustering: A review. ACM Computing Surveys, **31**(3), 264-323. doi:10.1145/331499.331504

[23]. Shirkhorshidi, A.S., Aghabozorgi, S., Wah, T.Y., & Herawan, T. (2014). Big Data Clustering: A Review. In Murgante B. et al. (eds), *International Conference on Computational Science and Its Applications* (Vol. **8583**, Lecture Notes in Computer Science, pp. 707-720). Springer.doi:10.1007/978-3-319-09156-3_49

[24]. Pujari, A.K., Rajesh, K., & Reddy, D.S. (2001). Clustering Techniques in Data Mining—A Survey. *IETE Journal of Research,* **47**(1-2): 19-28. doi:10.1080/03772063.2001.11416199

[25]. Dave, M., & Gianey, H. (2016). Different clustering algorithms for Big Data analytics: A review. In *Proceedings of IEEE International Conference System Modeling & Advancement in Research Trends* (pp. 328-333). *IEEE Xplore Digital Library*. doi:10.1109/sysmart.2016.7894544

[26]. MacQueen, J. (1967). Some Methods for Classification and Analysis of Multivariate Observations. *In Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, Vol. **1**, pp. 281-297.

[27]. Pandey, R., Mohan, L., Bisht, S. and Pant, J. (2017). Data Mining and Data Warehouse. *International Journal on Emerging Technologies,* (Special Issue NCETST-2017) **8**(1): 155-157.

[28]. Shrivastava, J. and Shrivastava, N. (2014). A Review of Data Reduction/ Extraction in Data mining from the Large set of Database. *International Journal of Electrical, Electronics and Computer Engineering,* **3**(2): 149-153.

[29]. Joshi, G. and Pandey, P. (2017). DATA MINING: A Comparative Study on Data Mining techniques. *International Journal on Emerging Technologies,* (Special Issue NCETST-2017) **8**(1): 373-376.

[30]. Khan, S.S. and. Quadri, S.M.K. (2016). Prediction of Angiographic Disease Status using Rule Based Data Mining Techniques. *Biological Forum – An International Journal,* **8**(2): 103-107.

How to cite this article: Pandey, K.K. and Shukla, D. (2019). A Study of Clustering Taxonomy for Big Data Mining with Optimized Clustering MapReduce Model. *International Journal of Emerging Technologies*, **10**(2): 226–234.