An Analysis of Different Clustering Techniques for Managing Big Data and Data Mining

Ms. Archana, Research Scholar, Department of Computer Science, Singhania University, Rajasthan.
Dr. Gaurav Aggarwal, Research Supervisor, Department of Computer Science, Singhania University, Rajasthan.

Abstract:

A Big Data is important field in the IT. It is highly trending due to the substantial advancement in use of internet. Clustering a data mining technique, has been used for the management of provided data in the medical and biology field. The cause is that the amount of data accumulated is high and not structured. The data has been gathered to the different physical and machine language form. For the management of these datasets, a lot of algorithms. Such algorithms that may be used are Map Reduce dependent on the K-mean clustering algorithm. Here datasets has been arranged and an entity has been developed for the determination of the similarity between objects. Map Reduce has been known as a potentially labeled calculating paradigm applied in large-scale execution of data in cloud computing. The deal with big data has been made for decrease the complexity and got the transparency. Research also focused on classification operation in area of healthcare that are based on machine learning.

Keywords:

Big Data, Data Mining, Clustering and Data processing .

1.1 INTRODUCTION:

The term "big data" was coined in the early 1990s. It has gained a following and a level of significance that is only going to rise in the coming years. Big data is now an essential part of every company's overall strategy. The MGI has articulated Big Data in the form of datasets, according to its report. There is no particular database program that can handle the scale of this data set. More information is being released into the world each and every day. Digital, social, and internet media use adds a layer of complexity and firepower. The rate at which new information may be gleaned is astoundingly quick these days. A variety of information is available since it comes from diverse sources and might be an important differentiator in today's competitive environment. Data from the same group or cluster are more similar than data from different groups or clusters. This is a major data issue. With big data, applications may be leveraged for anything from health care to marketing to city planning to seismic research to online document classification.

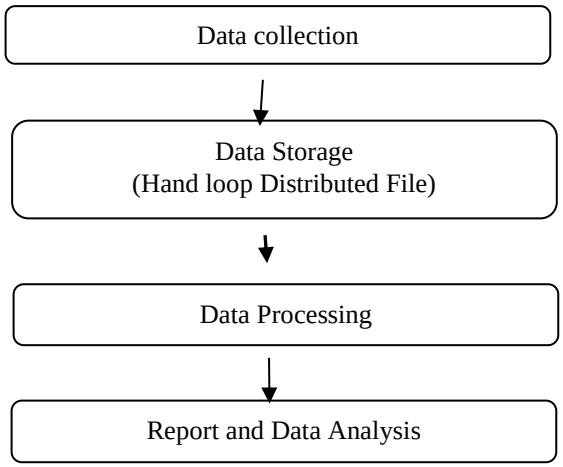

Fig 1 Big Data Processing

1.2 Data Processing

Data mining means that the removal of attractive, important, conversant data from huge databases. On the basis of present data it forecast the formation of group and metrics together and its association in order to identify its events. It occurs together or in a sequence. At last it detects the outliers not follows the required desired behavior.

1.2.1 Data Cleaning

The process to remove the noise as well as the inconsistency has been know as data cleaning, this data has been removed from the database. In this step, one tries to fill in the missing values, identify as well as to remove outliers. Here the inconsistencies are also resolved.

1.2.2 Data Integration

Data have been merged and generated in this stage for various servers. Some of the problems that may arise here include schema integration, redundancy, conflicts in data value detection and resolution.

1.2.3 Data Selection

Data selection is also essential step in the preprocessing of data. Data relevant to evaluation job has been achieved to the database.

1.2.4 Transformation of Data

Data has been transformed or consolidated into forms suitable to mining. Smoothing, Aggregation, Generalization are included in Data transformation. As well as the normalization as well as the attribute construction is considered here.

1.2.5 Data Mining

The process of it is analysis of data from different viewpoints. Process of scanning in order is using pattern recognition technology and statistical and mathematical methods to identify significant new connections, patterns and trends.

1.3 Clustering

Clustering is a process of data that is termed clustering in a collection of relevant subclasses. Used either as an autonomous tool for understanding the distribution of data or as a preprocessing step for other algorithms. Cluster analysis is job of combination items in cluster of the same kind. [8]

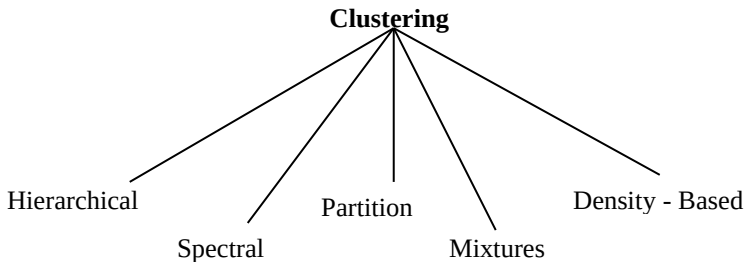

Fig 2 Clustering

1.3.1 Requirements of Clustering

Scalability:

To deal within large databases, highly scalable clustering algorithms are required.

High dimensionality:

The Clustering algorithms are able to handle low-dimensional data. Along with this such provides the high dimensional space.

Dealing with noise data – Databases includes the noise, missing etc. Some algorithms are not insensitive to this data and provide poor quality clustering.

Interpretability:

The results got after the Clustering, should be interpretable, comprehensible. It should be usable.

1.3.2 Clustering Methods

Clustering technique has been divided into below given categories:

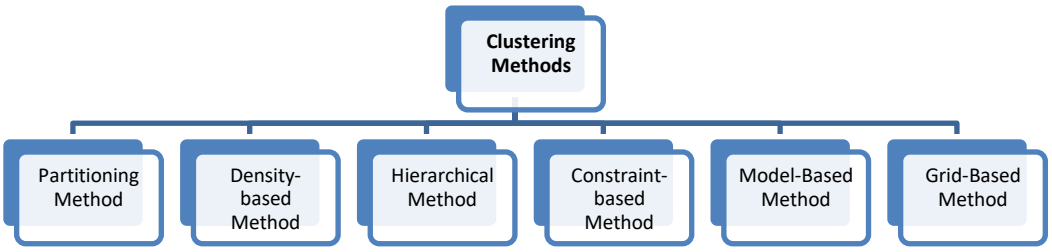

Fig 3 Clustering methods

1.4 Machine learning

Observation has shown that an industrial setup should be able to undertake predictive operations. That which is too sophisticated for machines may be learned by machine learning methods by recording data from several sensors that are linked to machines, and then discovering anomalous operations. The most widely used artificial intelligence technique is machine learning. An artificial intelligence (AI) is a specific intellect that is used by machines. The idea of incorporating the clustered notion has been put up in this article. Reinforcement learning environments may save time and space by using this method. A significant AI application is machine learning. As a result, the system is able to grow and develop. Experimenting and not being entirely programmed helps it grow. It has been discovered that machine learning tends to concentrate on creating computer software that can execute program and be used by a single user. To put it another way, AI is all about making machines capable of experiencing for themselves. Machine learning has long been acknowledged as a frequent application of artificial intelligence. One of the most prominent terms used to describe artificial intelligence is "AI," which stands for "Artificial Intelligence." Here, the clustering concept has been provided for consideration. This tool may save both time and space in reinforcement learn contexts. Machine learning is an important AI application. It allows the system to grow and improve. When it's not completely programmed, it's able to expand. Machine learning focuses on computer programs that can be run and utilized by a single human. The basic objective of artificial intelligence is to allow computers to experience things on their own, without the assistance of a human. These scenarios do not need human intervention. There are several benefits to using machine learning, including the following:

1. As well as retail and wholesale, it's found in the financial services industry as a whole as well as the healthcare industry in general.
2. Devices are able to reduce their time cycle, which improves their resource efficiency.
3. This strategy is often used by social media platforms like Google and Facebook to push advertisements depending on the preferences of its users.

4. Quality can be improved in big and intricate process systems using a variety of machine learning technologies.

1.5 LSTM

Hochreiter and Schmidhuber created LSTM networks in 1997, and they have since established accuracy records across a wide range of application sectors.

When LSTM started revolutionising speech recognition in 2007, it was able to outperform previous models in a variety of speech applications. For the first time in 2009, a CTC-trained LSTM network won three connected handwriting recognition contests. Baidu's CTC-trained RNNs beat the 2S09 Switchboard Hub5'00 speech recognition dataset benchmark in 2014 without using any typical speech processing methods.

Text-to-speech synthesis and large-vocabulary voice recognition were both enhanced using LSTM. Google's voice recognition is improved by 49% in 2015 due to CTC-trained LSTM. Language Modeling and Multilingual Language Processing were enhanced by LSTM **[2]**

1.6 Deep learning

Deep learning is one of the many sub domains of machine learning. It is able to learn from well-structured and manageable data and provide valuable and knowledge-rich output. Deep neural learning is a term for this kind of learning. The term "deep neural network" was coined by a single researcher. Algorithms fall within the topic of machine learning. Large amount of data is taken into consideration in deep learning. An investigation has shown that machine learning systems must be supervised. The programmer is responsible for defining his instructions. For example, if someone wants to find a picture of a dog, he must describe his search to the computer in great detail. Autonomous software with an extensive knowledge base produces better results for every search. Simple machine learning isn't as good as deep learning since it's slower and less precise.

It's a well-known fact that learning may be very effectively structured. Another term for it is "hierarchical learning." Machine learning methods are being investigated in a bigger context. These methods rely heavily on artificial neural networks.

2. Literature Review

Hassan Ismail Fawaz, 2019 [11] studied by doing an empirical investigation of the most recent DNN designs for TSC. The most effective deep learning applications in diverse time series domains are summarised under a single taxonomy of DNNs for TSC. Open source deep learning framework is made available to the TSC community, where we deployed each of the comparison algorithms, as well as assessed them on the UCR/UEA archive and 12 multivariate time series datasets. We provide the most comprehensive research of DNNs for TSC to date, by training 8730 models on 97 time series datasets.

Chen, G., & Islam, M. (2019). [5] Implemented big data analytics for healthcare. Big data analytics has been applied to aid process of care delivery along with exploration of disease. Author considered the utilization of huge volumes of data set that is related to medical field.

In 2018, Andreas Holzinger et al. [2] were at the beginning of a new AI era. In spite of all of this, there are still a number of issues that need to be addressed, including the fact that the highest performing models are considered black boxes.

In 2018, Inés Sittón Candanedo et al. [12] would be the focus of this study for Using of machine learning methods. Predictive models for the Industry 4.0 environment will be developed utilising machine learning methods and the aforementioned dataset, as detailed in this article.

Kaur, J., Sachdeva, K., & Singh, G. (2018). [13] presented the Image processing on multimode Hadoop cluster. Research represents that data produced over internet is increasing day by day at exponential rate. The classical methods are insufficient for processing is termed as 'Big Data'. There are various tools in Hadoop to analyze the textual data such as Pig, base, etc.

Mohammed, A. Q., & Bharati, R. (2018). [16] Proposed research in order to improve resource utilization during big data processing in heterogeneous network. Research has make use of Hadoop MapReduce in order to achieve objective. Simulation concludes that there is increment in resources utilization and decrement in job running time by 10-30 %

Merla, P. R., & Liang, Y. (2017). [17] presented the Data analysis using Hadoop map reduce environment. The research work has provided the evaluation of YouTube data. They have done this with the use of Hadoop map reduce system on the base of cloud platform AWS.

Mohammad Hesam Hesamian, 2018, [18] gave a thorough evaluation of common algorithms for medical picture segmentation that use deep-learning techniques. In addition, they provide an overview of the most typical problems encountered and their remedies.

Bhardwaj, A., Singh, V. K., Vanraj, & Narayan, Y. (2016). [3] did research on Analyzing Big Data with Hadoop cluster in hd insight azure Cloud. Presently Cloud dependent Hadoop is gaining huge interest. It is offering ready to use Hadoop cluster environment in order to process Big Data. It has been excluding the operational issues in case of on-site hardware investment, IT support, and installing, configuring of Hadoop components such as HDFS and Map Reduce.

Com, K. M. R. B. (2016). [6] has explained the method of data mining. A cluster of data objects is used as a group. To do the cluster evaluation, first partition is to arrange the data into groups. It is done on the base of data similarity **Bhandarkar, M. (2010). [4]** wrote on Map Reduce programming with apache Hadoop. The researcher has explained the solutions of general issues that are encountered in maximizing Hadoop application performance.

.**Diwate, R. B., & Sahu, A. (2014). [7]** This research on data mining techniques. Their research has provided a feature classification technique using association rule. The major objective of research has to consider various data mining techniques. Research has considered percentage of the most widely data mining techniques which are found useful for normal life.

[3] Problem Statement

There have been several researches in field of big data and health care. It has been observed that previous research did limited work in clustering. Moreover these researches have faced performance and accuracy issues. It is observed that previous research need to be more scalable, flexible and efficient. In other words Clustering research in the past has been found to be sparse. Furthermore, these investigations have run into problems with speed and precision. Previous studies have shown a need for increased scalability, flexibility, and efficiency. The present problem entitled "**Managing Big Data using Enhanced Clustering Mechanism**", firstly arrives due the lack of accurate clustering technique as data has been gathered in huge volume. It is coming from un-trusted sources on different websites. These systems are using various clustering methods to handle the BIG Data. Due to ignorant behavior of medical teams people are suffering from several side effects that are caused due to prescribed medicine.

Table Comparative analysis of Features

Citation	Clustering	Big Data	Multinode Hadoop Cluster	Data Mining	Machine Learning	Deep Learning
[1]	No	No	No	Yes	No	No
[2]	No	No	No	No	Yes	No
[3]	Yes	Yes	Yes	No	No	No
[4]	No	No	Yes	Yes	No	No
[5]	No	Yes	No	No	No	No
[6]	No	No	No	Yes	No	No
[7]	No	No	No	Yes	No	No
[8]	No	No	No	Yes	No	No
[9]	Yes	No	No	No	No	No
[10]	No	No	No	Yes	No	Yes
[11]	No	No	No	No	Yes	No
[12]	No	No	Yes	No	No	No
[13]	Yes	Yes	No	No	No	No
[14]	No	No	No	Yes	No	No
[15]	Yes	Yes	Yes	No	No	No
[16]	No	No	Yes	No	No	No
[17]	Yes	No	Yes	No	No	No

It has been observed that more than 70% of health related issues exists due to lack of nutrition. Also, the clustering mechanism used in various research papers does not provide the efficient real time processing in order to resolve the queries of the patient. Thus there is need to introduce a Nutrition prescription system that would focus on nutritional deficiency. Moreover there is need to improve accuracy and performance of AI based classification process.

[4] Proposed work

In proposed model healthcare Dataset has been considered for training and Attributes are eliminated consider ship algorithm. Some attributes are eliminated that have single value in all cases. Data set after filtering the feature selection mechanism is applied. Then, 70% of the data is used for training, and 30% is used to test. Finally, each of the following layers is applied: the LSTM, fully linked, and soft max layers. Predicting the onset of illness necessitates the use of classification. Once the predicted value has been obtained, a confusion matrix is generated that takes into account both the expected and actual values in order to arrive at the four possible outcomes: true positive, false positive, true negative, and false negative. Overall accuracy is determined by calculating the accuracy, precision, recall, and fscore of the results. This section has presented the issues in existing researches with proposed work. Then the objectives of research are presented. LSTM mechanism applied in research has been explained and the two model developed in proposed implementation are also discussed. Data set used for training the categories, sub categories used in research are explained here.Proposed work has made use of LSTM mechanism. Two separate models of the LSTM have been used to create two distinct trained networks. The first model uses a single LSTM layer, whereas the second model uses two LSTM layers and a drop-out layer to get better results.

Algorithm for model 1

1. Get the dataset
2. Select the features in order to train the dataset
3. Set the ratio of training to 70% and testing to 30%.
4. Apply LSTM layer
5. Apply fully connected layer
6. Apply softmax layer
7. Perform classification
8. Perform decision making for classification.

Firstly two parallel processes are there to get the keywords and queries respectively from the health care center dataset. The map reduce process is then applied on the obtained keywords and queries in order to obtain the frequency. The get frequency process is used to draw the frequency of the keywords and queries after the map

reduce process. The frequency data set is used to store the obtained frequency. After that clustering mechanism is followed to make the clusters.

Dataset:

These are the centralized location where information regarding cure and cause is stored in unstructured format. The centralized dataset holds the whole information about the disease and their symptoms, the deficiency of which factor causing that disease, but all this information is stored in unstructured manner.

Queries:

Query is the portion of dataset where the views of patient and doctors are considered. When dealing with an unstructured dataset, the map-reduce algorithm is used to determine the frequency of a certain keyword in a query.

Map Reduce:

Mechanism is to obtain keyword frequency in the dataset. Map Reduce programmes may run large datasets. It is done using low-cost computers called clusters. In general, individual systems are characterized as nodes in a cluster. Map Reduce covers the two major phases of computing. It was applied sequentially to large quantities of data.

Enhanced K Mean clustering:

K mean is simplest unmonitored learning methods to tackle famous issue of clustering. Method follows straightforward & quick approach to classify certain data collection. It is done via a no. of clusters which assume k clusters with a fixed apriority. Main key is k centres to specify. These centres must be positioned in an intelligent way. There are various places where the results are varied. Therefore, it is preferable to distance them from one other. Another stage is to examine each item that is part of a certain data collection. Then the closest centre is linked. If there is no point left outstanding, the first stage is finished. This is the method to get early group age. At this time k new centroids must be re-calculated as a baricenter of clusters arising from tranditional stage. Once new bond must be made b/w similar data sets & closest new centre. There is a loop formed. Due to this loop it may be seen that k centres progressively change their position. It will be done until changes are made or, in other words, centres no longer move.

Optimization:

In order to take decision there is need of optimized data that has been retrieved from K mean clustering module. Prescription logic is used on clustered data, which is the important information gleaned from the whole dataset, to arrive at a choice.

Result and Discussion

The proposed work has been found significant for the processing of the big data. The proposed work has categorized the contents in order to increase the performance. The content retrieval has been found fast as compare to traditional work. The MATLAB simulation represents that the keywords have been clustered according to their frequency. In this way it is clear to the system, how much content are available corresponding to particular keyword. The symptoms have been mapped to the nutrition keywords. These nutrition keywords are mapped to the source and description. The MATLAB simulation also represented the dynamic curve that is showing the performance comparison of traditional & current work. Map reduce mechanism has played significant role in extracting the contents frequency from the given dataset. Proposed work has provided more efficient, accurate and high performance solution.

The most popular method of grouping data is to break it down into smaller groups that are more closely linked. Data may be clustered before running a learning algorithm, or it can be used as a statistical tool to identify interesting patterns in the data. The clustering technique is used in a range of applications, such as market research, customer segmentation, biomedical imaging, search result clustering, recommendation engines and pattern recognition, as well as image processing. Clustering is a typical strategy used by retail firms to identify similar groups of families.

1. **Partitioning-based Clustering**

Data is divided into k clusters by partitioning the objects, with each partition forming/representing a single cluster. Each cluster comprises at least one data item, and each data object must be allocated to a single cluster.

2. **Hierarchical-based Clustering**

Clusters with a tree-like structure are produced applying these clustering techniques depending on the hierarchical structure of the data.

3. **Density-based Clustering**

These clustering algorithms discover dense regions with some similarity and separate them from the less dense portions of the space they are compared to. The accuracy and combinatorial power of these technologies make it possible to integrate two clusters.

4. **Grid-based Clustering**

To create a grid structure, data is grouped into a limited number of cells using a manner that resembles a grid. Fast and non-dependent on the quantity of data items at any one moment make these grids ideal for clustering operations.

5. **Model-based Clustering**

According to these methods, a mathematical model is used to fit and then optimize the data. Conventional statistics is then used to determine how many groups there are.

The Comparative chart of cluster size has been simulated using MATLAB. After getting the frequency of cluster 1, cluster 2, cluster the data is graphically simulated using MATLAB script. The x coordinate is representing the cluster. The y axis is representing the maximum frequency of keyword in corresponding cluster. Research has provided more flexible, scalable and efficient solution.

Simulation of classification

The training process of proposed model is shown below. Here the dataset has been trained using LSTM in order to classify considering trained set. The validation accuracy is 89.73% in this case. The accuracy is influenced by LSTM layer, dataset, number of hidden layers and batch size.

[6] Conclusion & Future Scope:

The management of big data at health care center is done through the proper clustering mechanism. It has been concluded that after clustering of complete unstructured healthcare dataset. The material has been broken down into tiny databases and the patient inquiries have been simulated in real time. The amount of time it takes to get a prescription has also dropped. Checking the file size of each cluster yielded the cluster's size. Cluster 1 has a higher keyword density than clusters 2 and 3, according to the most recent analysis. MATLAB has been used to simulate a cluster size comparison chart. MATLAB script is used to create a graphical representation of the data once the frequency of clusters 1, 2, and 3 has been calculated. X coordinate is representing cluster. The y axis is representing the maximum frequency of keyword in corresponding cluster. Research has provided more flexible, scalable and efficient solution.

Health care centers consist of huge data sets. Therefore it is necessary to classify this data using best clustering mechanisms. In the proposed research, sets of big data have been stored on a centralized server. After that, the data has been shared between multiple clients. Map Reduce technique is capable to be used for faster data access. It would be helpful for patients from different locations. In order to use this method, patients will be able to connect to a particular and specialized health care facility. In this way they can attain the better services as per the requirement. As there are the facilities of backup in such system, there would be secure access for patient. In future the keywords and contents length might increase.

References :

[1] Amato, F., Cozzolino, G., Moscato, F., Moscato, V., Picariello, A., & Sperli, G. (2019). Data mining in social network. *Smart Innovation, Systems and Technologies*, *98*, 53–63. https://doi.org/10.1007/978-3-319-92231-7_6

[2] A. Holzinger, P. Kieseberg, E. Weippl, and A. M. Tjoa. (2018). Current advances, trends and challenges of machine learning and knowledge extraction: From machine learning to explainable AI, vol. 11015 LNCS. Springer International Publishing.

[3] Bhardwaj, A., Singh, V. K., Vanraj, & Narayan, Y. (2016). Analyzing BigData with Hadoop cluster in HDInsight azure Cloud. *12th IEEE International Conference Electronics, Energy, Environment, Communication, Computer, Control: (E3-C3), INDICON 2015*. https://doi.org/10.1109/INDICON.2015.7443472

[4] Bhandarkar, M. (2010). *MapReduce programming with apache Hadoop*. 1–1. https://doi.org/10.1109/ipdps.2010.5470377

[5] Chen, G., & Islam, M. (2019). Big Data Analytics in Healthcare. *Proceedings - 2019 2nd International Conference on Safety Produce Informatization, IICSPI 2019, 2015*, 227–230. https://doi.org/10.1109/IICSPI48186.2019.9095872

[6] Com, K. M. R. B. (2016). Data mining techniques. *SpringerBriefs in Applied Sciences and Technology*, *179*(10), 13–30. https://doi.org/10.1007/978-3-319-22294-3_3

[7] Diwate, R. B., & Sahu, A. (2014). Data Mining Techniques in Association Rule: A Review. *International Journal of Computer Science & Information Technology*, *5*(1), 227–229. http://connection.ebscohost.com/c/articles/99091355/data-mining-techniques-association-rule-review

[8] Gupta, M. K., & Chandra, P. (2020). A comprehensive survey of data mining. *International Journal of Information Technology (Singapore)*, *12*(4), 1243–1257. https://doi.org/10.1007/s41870-020-00427-7

[9] Gupta, M. K., & Chandra, P. (2019). MP-K-Means: Modified Partition Based Cluster Initialization Method for K-Means Algorithm. *International Journal of Recent Technology and Engineering*, *8*(4), 1140–1148. https://doi.org/10.35940/ijrte.d6837.118419

[10] H. Ismail Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P. A. Muller. (2019). Deep learning for time series classification: a review. Data Min. Knowl. Discov., vol. 33, no. 4, pp. 917–963, 2019. doi: 10.1007/s10618-019-00619-1.

[11] I. S. Candanedo, E. H. Nieves, S. R. González, M. T. S. Martín, and A. G. Briones (2018). Machine learning predictive model for industry 4.0," Commun. Comput. Inf. Sci., vol. 877, pp. 501–510, 2018, doi: 10.1007/978-3-319-95204-8_42.

[12] Kaur, J., Sachdeva, K., & Singh, G. (2018). Image processing on multinode hadoop cluster. *International Conference on Electrical, Electronics, Communication Computer Technologies and Optimization Techniques, ICEECCOT 2017*, *2018-January*, 21–26. https://doi.org/10.1109/ICEECCOT.2017.8284515

[13] Kumar D. (2015). A Hybrid Approach to Clustering in Big Data.

[14] Lin, C. Y., Yu, K. M., Ouyang, W., & Zhou, J. (2011). An OpenCL candidate slicing frequent pattern mining algorithm on graphic processing units. *Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics*, 2344–2349. https://doi.org/10.1109/ICSMC.2011.6084028

[15] Lathiya P. (2016). Improved CURE clustering for big data using Hadoop and Mapreduce.

[16] Mohammed, A. Q., & Bharati, R. (2018). An efficient technique to improve resources utilization for hadoop MapReduce in heterogeneous system. *ICCT 2017 - International Conference on Intelligent Communication and Computational Techniques*, *2018-January*, 12–16. https://doi.org/10.1109/INTELCCT.2017.8324012