I.J. Education and Management Engineering, **2018, 4, 61-70**
Published Online July 2018 in MECS (http://www.mecs-press.net)
DOI: 10.5815/ijeme.2018.04.06

Available online at http://www.mecs-press.net/ijeme

An Intelligent Distributed K-means Algorithm over Cloudera /Hadoop

Tawseef Ayoub Shaikh[a], Umar Badr Shafeeque[b], Maksud Ahamad[c]

Department of Computer Engineering, Aligarh Muslim University, Uttar Pradesh, India

Received: 29 January 2018; Accepted: 16 April 2018; Published: 08 July 2018

Abstract

The 21st century evolved with tsunami of data generation by the human civilization that has delivered new words like Big Data to the world of vocabulary. Digitization process has almost overtaken all the major sectors and it has played a pivotal role of dominance as for as virtual digital world is concerned. This in turn has landed us in most debated term "Big Data" in the present decade. Big Data has made the traditional relational databases (RDMS) handicapped in terms of their huge size and speed of its creation. The hunger to manage and process this gigantic complex heterogeneous data, has again followed the age old rule of "Necessity is the mother of Invention", and came up with idea of HadoopMapReduce for the same. The given work uses K-Means clustering algorithm on a benchmark MRI dataset from OASIS database, in order to cluster the data based upon their visual similarity, using WEKA. Until a threshold size it worked out and after that compelled WEKA to prompt an emergency message "out of memory" on display. A Map/Reduce version of K-means is implemented on top of Hadoop using R, so as to cure this problem. The given algorithm is evaluated using Speedup, Scale up and Size up parameters and it neatly performed better as the size of the input data gets increased.

Index Terms: Big data, Healthcare informatics, MRI (Magnetic Imaging Resonance), Clustering.

1. Introduction

With the improving evolutions in the science and related research, humanity has always been at the risk of vulnerable diseases threatening the human wellbeing. Recent ally, the field related to brain diseases/disorders got illuminated which was almost ignored because of the lack of knowledge and information. For quenching thirst of the same, scientists with their tier efforts come up with more reliable and sophisticated tools aiming at

* Corresponding author. Tel.: +91-9622513326
E-mail address: tawseef37@gmail.com[a], shafeeque.umar@gmail.com[b], maksud.ahmad12@gmail.com[c]

easy and precise diagnosis of the related problems. Magnetic resonance imaging (MRI), fMRI (Functional Magnetic resonance imaging (MRI)), CT (Computed Tomography) Scans, EEG (Electroencephalography) etc. are such techniques known to everyone nowadays.

For fabricating in depth images of the inside of the body, MRI scan equipped with waves and strong magnetic fields is worth use of which contains large ample space to lay down a patient inside the tube during the scan [1]. Nearly any part of the body can be scrutinised by an MRI scan, including the liver, blood vessels, brain and spinal cord, bones and joints, breast, heart, internal organs, such as womb or prostate gland. Various nomenclature of MRI's does exist depending on which part of body the MRI is carried on. Some of them functional in modern clinical investigation are Cardiac MRI, Chest MRI , Knee MRI, Magnetic Resonance (MR) Defecography [2], Magnetic Resonance (MR)-Guided Breast Biopsy, Magnetic Resonance Cholangiopancreatography (MRCP), Magnetic Resonance Imaging (MRI) of Spine, Magnetic Resonance Imaging (MRI) of Head , Magnetic Resonance Imaging (MRI) of Dynamic Pelvic Floor, Magnetic Resonance Functional (fMRI) of Brain, MR Angiography (MRA), MR Enterography, MRI of the Body (Chest, Abdomen, Pelvis), MRI of Musculoskeletal System, MRI of the Prostate, Shoulder MRI etc. [3].

Metabolic changes taking place inside the brain are measured by functional magnetic resonance imaging (fMRI) [3]. Inspection of the brain's anatomy, govern brain parts liable for conducting critical functions, weighing effects of stroke or disease are the outcomes from the fMRI. Oddities indoors the brain that remain invisible with other imaging techniques are unearthed by it. In addition in explaining brain anatomy, MRI aids Neurosurgeons also in assessing integrity of the spinal cord after trauma. It serves in the scenarios also when problems associated with the vertebrae or intervertebral discs of the spine are bored in mind. Shape and structure of the heart and aorta (so as to detect aneurysms or tears) also part of its offerings.

Outcomes of an MRI scan can aid in diagnose conditions, plan treatments and weighing the success of prior dealing [4]. Accuracy of disease uncovering throughout the body is at the core of MRI techniques and frequently comes to rescue after former testing fails in delivering abundant evidence for sanctioning a patient's diagnosis. While Bleeding or swelling in head signals trauma to the brain, other deviations regularly mined embrace brain tumors, stroke, aneurysms as well as tumors or inflammation of the spine. In addition in enlightening about the information on glands and organs within the abdomen, information about soft tissues, joints structure and bones of the body is also offered by it. These fruitful services can in turn assist in deferring surgery or more exactly focused after expressing results of an MRI scan [5].

Remaining paper is fashioned as: Chapter 2nd throws a torch on the related work done in the past in the same direction. Chapter 3rd concentrates on the description of the MRI dataset from OASIS database, used in this study. It also throws a brief light on the methodology of extracting the productive feature vector from large high dimensional feature space and clustering of the dataset on Weka tool. Experimental results are carried and plotted out in graphical in Chapter 4th and the results are discussed and debated in Chapter 5th in discussion part and finally the Conclusion of the findings of this study is concluded in Chapter 6th.

2. Related Works

This Chapter revolves around the work already carried out in the said field providing a baseline about the new possibilities in which same/related work will be proceeded further.

Addressing the problem for a deep web data source, Tantan Liu and Gagan Agrawal [6] came up with notion of a fresh stratified clustering method. For catching relationship between input and output attributes, stratified k-means clustering method is offered in their work which stratifies space of input attributes of a deep web data source. Achieving an optimal appraisal for statistics, including proportions and centers within sub-spaces of the output attributes, corresponding space of output attributes of deep web data source is separated into sub-spaces. Two synthetic and two real datasets are used for gauging their method and results offered substantial gains in assessment of accuracy from both the novel aspects of their work i.e., the use of stratification (5%-55%) and representative sampling methods (up to 54%).

Immovability of k-means clustering is fully dependent on the number of optimal solutions for underlying

optimization problem in data distribution, got focused from pioneering work from the authors of work in [7]. The results of this work encountered the public faith and practice which views stability as a validity pointer, or meaningfulness, of the choice of a clustering algorithm plus number of clusters.

Yujie Xu, Wenyu Qu, Zhiyang Li, Geyong Min, Keqiu Li and Zhaobin Liu [8] contributed in their work by coming up with an innovation for resourcefully instigating traversals of large-scale RDF graphs over MapReduce, grounded on Breadth First Search (BFS) strategy for visiting (RDF) graphs. Promised optimal speeds-up in the analysis of RDF graphs with respect to competitor approaches are demonstrated in the implementation work.

In [9], Yingchi Mao, Ziyang Xu, Ping Ping and Longbao Wang established a Hadoop distributed cluster built on the Cloud Stack and realized the optimal distributed K-Means clustering algorithm centered on Map/Reduce. In addition to execution time efficiency, their approach confirms to be a good candidate for best quality of the clustering result in dealing with large-scale data set, which they confirmed from the experiment results.

For the solution of privacy-preserving multi-party k-means clustering problem, when data is vertically partitioned and horizontally partitioned respectively among different parties, Teng-Kai Yu, D.T. Lee, Shih-Ming Chang and Justin Zhan [10] applied the notion of parallel computing. They offered algorithms for solving problems for these data partition models enjoying execution time of O (nk) and O (m(k + log(n=k))) respectively. The time complexities of the algorithms are far better than others where no parallel computing is incorporated.

MapReduce founded parallel k -means clustering algorithm is suggested by Weizhong Zhao, Huifanf Ma and Qing He in [11]. Proposed algorithm got majority votes in terms of proficiently scaling well over big datasets on even on commodity hardware.

3. Methodology Adopted

This section has its roots in presenting details about the MRI dataset and its various features and the feature engineering techniques fruitful for extracting productive feature vector from MRI raw data.

Magnetic Resonance Data (MRI) used in the present study is taken from publically available database from OASIS database available at http://www.oasis-brains.org [12, 13]. It contains data of the brain MRI images. The invaluable tireless efforts concreted this idea into its final shape came up from Dr Randy Buckner at the Harvard Hughes Medical Institute (HHMI) at Harvard University, the Neuro Informatics Research Group (NRG) at Washington University, school of medicine and Biomedical Informatics Research Network (BIRN), Washington University Alzheimer's disease research center, whose priceless dedication made this dream come true and made OASIS datasets freely available by [14]. The Open Access Series of imaging Studies (OASIS) is a project aiming at making datasets of the brains freely available to the scientific community whose sole idea and hope is to make future discoveries in clinical and neuroscience which will provide an innovative platform and breakthrough in automation of the diagnosis/prediction of the Neurogenic diseases/brain disorders using the novel machine learning concepts, which in turn will make the healthcare optimal and qualitative [15].

The work we carried out got its initial shape by firstly downloading MRI medical images. Below is the stepwise procedure of the actual progress of this work.

Steps:

1) First of all, we downloaded a large number of MRI medical images from http://www.oasis-brains.org .
2) Then after pre-processing of the raw data, the corresponding histogram of each image was framed out in MATLAB and represented same by using a vector of (16*16*16) i.e. 4096 values in each single vector.
3) Then after .arff (attribute relation file format) [16] file format is generated.
4) Since there are 1000,000 histograms and 4096 values in each row thus making overall file size touching up to 34GB.

Finally after pre-processing and extracting the optimal feature vector from this ultra-feature space, our goal is to group (cluster) images according to their visual similarity. In order to carry out the same we implemented the services of the all-time Open source data mining tool Weka for the same. Till a particular threshold value of the input data size (approx. 30MB), Weka's Simple K-Means clustering groups the data in appropriate clusters using Euclidean distance as visible below in Figure 1 and 2 below. But when the data size is increased from 1000 instances to 100000 instances, it halts the Weka and makes it stuck outputting the message "Reading a File".

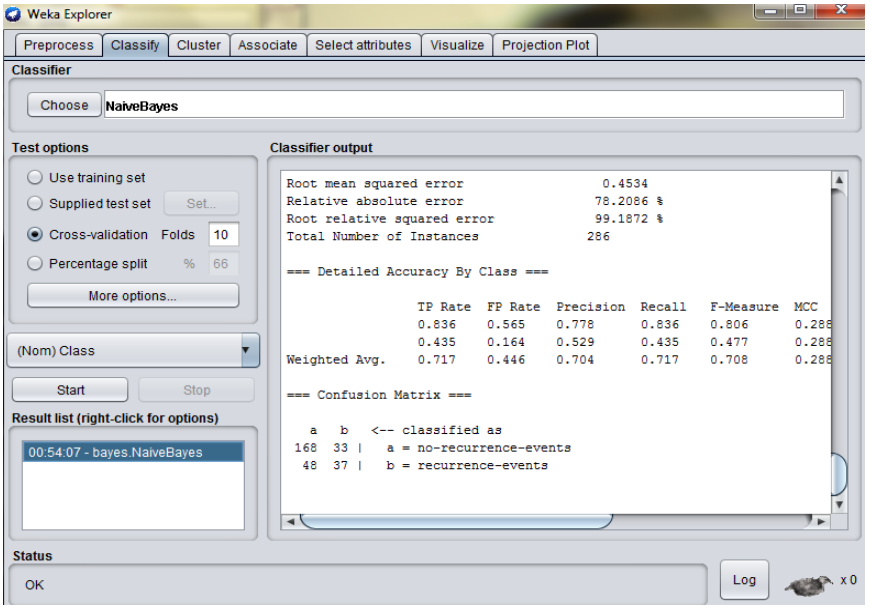


Fig.1. Screen shot of K-Means in Weka

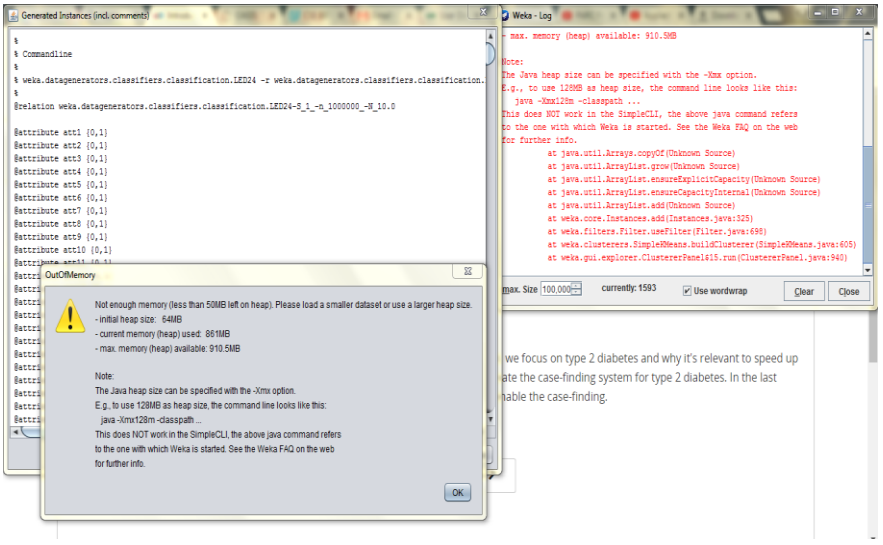


Fig.2. Screen shot of K-Means in Weka promoting message "Reading a File"

4. Empirical Results

This Chapter deals with the brief definition of Hadoop/MapReduce framework used in this work in addition to the metrics which are used as a measuring parameter for calculating the Scale-up, Size-up and Speed-up of the K-Means clustering over the Hadoop platform.

A. Hadoop/MapReduce:

So to solve the above issue we implemented the Parallel K-Means Clustering algorithm [17] using the Map/Reduce on top of Hadoop Cloudera environment [18] and implementing the same using 'R' language in coordination with the library RHadoop. Its working is sketched in the Figure 3 given below.

B. RHadoop contains three main R packages [19]:

1) Rhdfs: R interface for providing the HDFS usability from the R console.
2) Rmr: R interface for providing Hadoop MapReduce facility.
3) Rhbase: R interface for operating the Hadoop, Hbase data source.

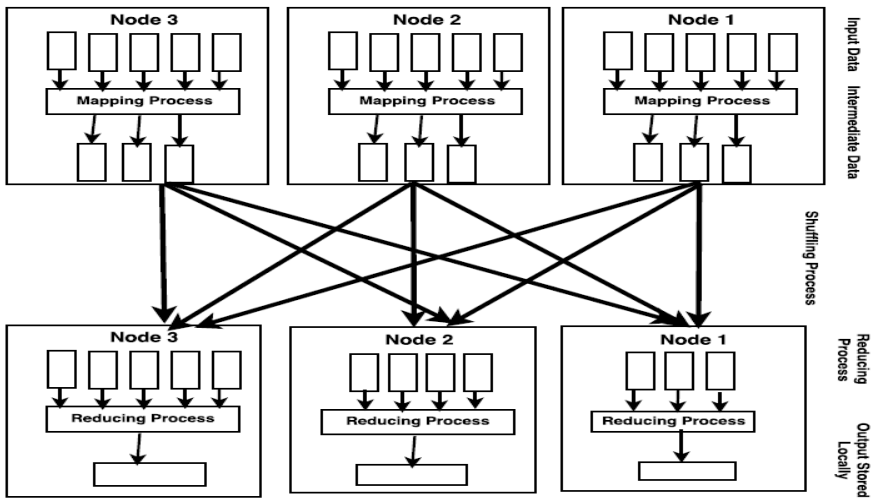


Fig.3. Working of Hadoop/MapReduce Framework

C. Installing R packages to connect R with Hadoop [20]:

1) rJava
2) RJSONIO
3) Itertools
4) Digest
5) Rcpp
6) Httr
7) Functional
8) Devtools
9) Plyr
10) Reshape2

Results Using K-means Clustering Hadoop MapReduce on 100MB dataset to which Weka got hanged. Clearly as the number of nodes in the commodity cluster got increased the CPU time accordingly gets decreased.

D. Hardware Configuration of our System:

We used the Cloudera's Quick start VM 5.4.2.0 Hadoop on 1, 2, 3 systems. Each of the system is having the following hardware configuration:

1) RAM of 8GB
2) 64 Bit Operating System
3) Intel Core i5-4590 with 3.30GHz
4) CentOS 6.4

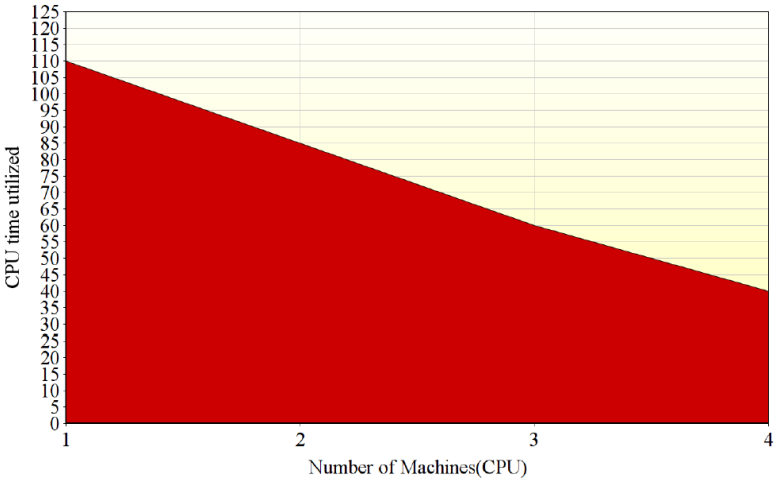
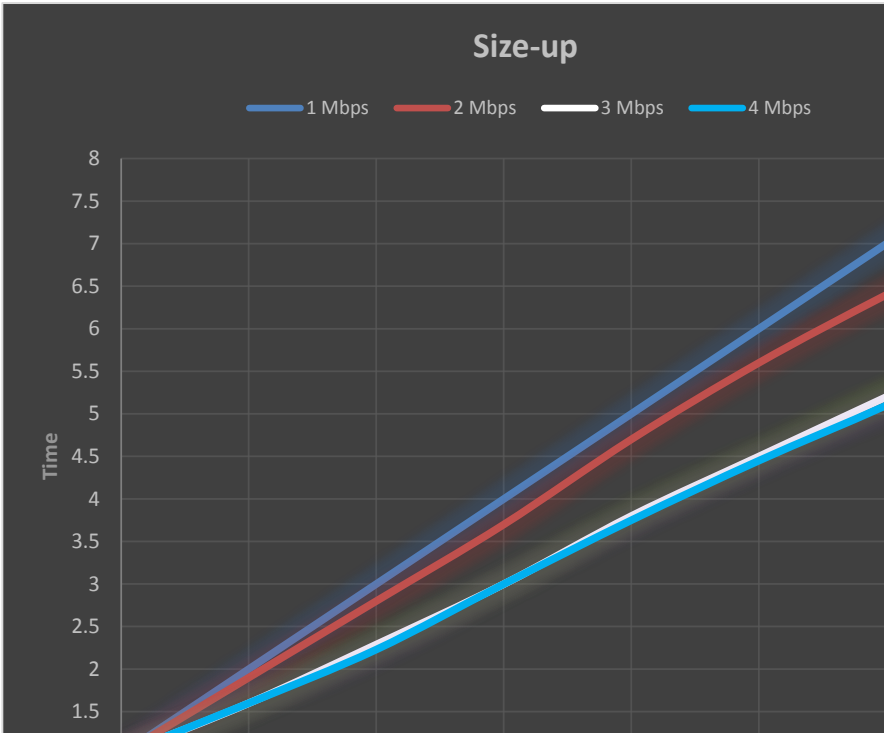


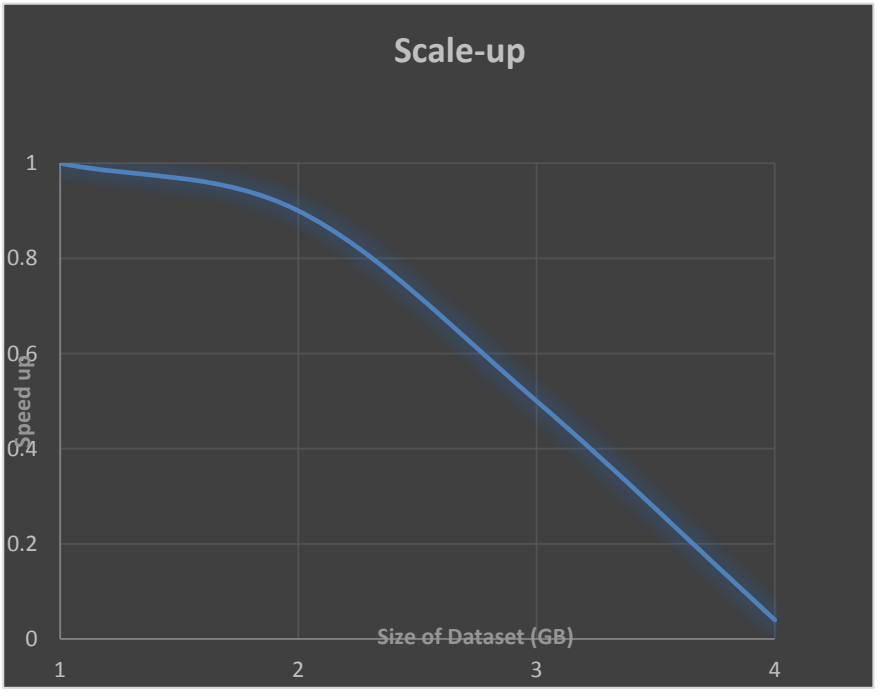
Fig.4. Number of Nodes vs CPU Execution Time

Number of machines/computers in the system is amplified and at same time constancy is maintained in dataset size, during measuring the speedup. Linear speedup, where a system with *t* times the number of computers harvests a speedup of *t*, was finally gained in the demonstration of perfect parallel algorithm. Speedup calculation on datasets possessing diverse sizes and systems are trademark of this study. The number of computers speckled from 1 to 3. Similarly, dataset size upsurges from 1GB to 8GB. The speedup for different datasets are showed in Fig.5. (a). Parallel K-Means has a very good speedup performance depicted from the results. Alternatively, as dataset size rises, performance of speedup gets better.

Algorithmic ability to cultivate both the system and dataset size, gives birth to the term Scaleup. It is defined as the ability of a *t*-times larger system to perform a *t*-times larger job in the same run-time as the original system. We have enlarged the datasets size in direct fraction to the number of computers in the system in this setup. Dataset size of 1GB, 2GB, 3GB and 4GB are executed on 1, 2 and 3 computers respectively in the scaleup experiment segment. The performance results of the datasets are depicted in Fig.5. (b).

Finally, we did the Sizeup evaluation of our proposed system. Datasets size grows by the factor *t* during encountering this phase while as number of computers in the system remain constant. Sizeup measures how much longer it takes on a given system, when the dataset size is *t*-times larger than the original dataset. We have fixed the number of computers to 1, 2 and 3 so as to measure the performance of sizeup, respectively. Fig.5. (c) shows the sizeup results on different computers.


Size-up
1 Mbps
2 Mbps
3 Mbps
4 Mbps
Time
8
7.5
7
6.5
6
5.5
5
4.5
4
3.5
3
2.5
2
1.5


Scale-up
Speed up
1
0.8
0.6
0.4
0.2
0
Size of Dataset (GB)
1
2
3
4

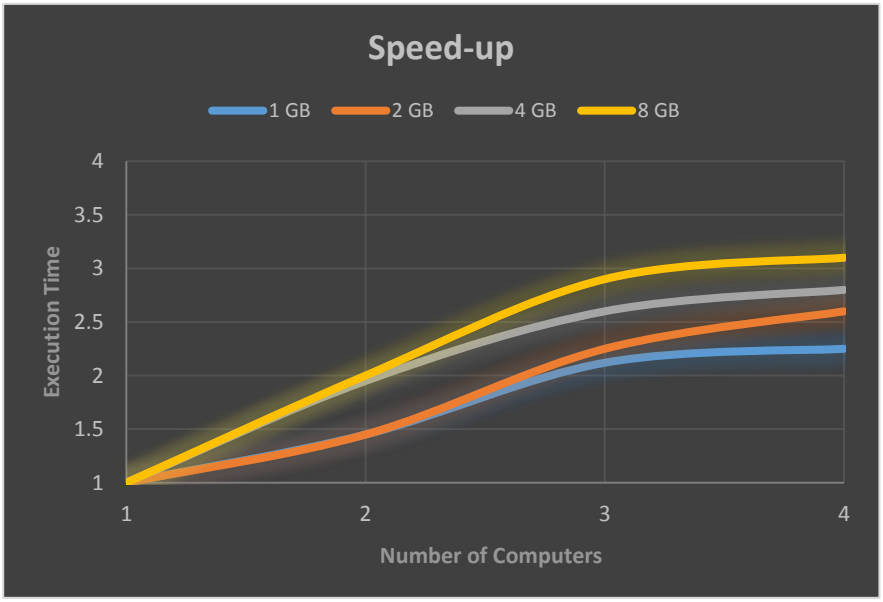


Fig.5. Evaluations results (a) Size-up (b) Scale-up (c) Speed-up

5. Discussions

The given work uses K-Means clustering algorithm on an MRI dataset from OASIS database, in order to cluster the data based upon their visual similarity by using WEKA. Until a threshold size of the input data, it worked out and after that compelled WEKA to prompt emergency message "out of memory". A Map/Reduce version of K-means is implemented on top of Cloud era Hadoop using R, so as to cure the problem arose out of the WEKA. The given algorithm is evaluated using Speedup, Scale up and Size up parameters and it neatly performed better as the size of the inputted data gets increased.

Speedup, Scaleup and Sizeup evaluation on datasets possessing dissimilar sizes and systems is at the heart of the present work. The number of computers speckled from 1 to 3 and dataset size upsurges from 1GB to 8GB during the speedup phase. Parallel K-Means has a very good speedup performance depicted from the results. Alternatively, as dataset size rises, performance of speedup gets better. Scaleup is defined as the ability of a *t*-times larger system to perform a *t*-times larger job in the same run-time as the original system. We have enlarged the datasets size in direct fraction to the number of computers in the system in this setup. Dataset size of 1GB, 2GB, 3GB and 4GB are executed on 1, 2 and 3 computers and the outcomes showed promising results. Moving forward same way, datasets size grows by the factor *t* while as number of computers in the system remain constant during. Sizeup phase is defined as the measure of how much longer it takes on a given system, when the dataset size is *t*-times larger than the original dataset. We have fixed the number of computers to 1, 2 and 3 so as to measure the performance of sizeup and results are promising as visible in Chapter 4^{th}.

The future is still open for expanding the dimensions of this work so as to come up with Hadoop/MapReduce versions of more and more machine learning algorithms particularly used in the day to day life of the common individual such as Recommendation Systems in Online Shopping Systems etc.

References

[1] Magnetic Resonance, Functional (fMRI) – Brain, Mar-16-2016, pp: 1-8. [https://www.radiologyinfo.org/en/pdf/fmribrain.pdf. [Last visited 5/11/2017].

[2] Savitz JB, Rauch SL, Drevets WC, Clinical application of brain imaging for the diagnosis of mood Disorders: the current state of play, *Molecular Psychiatry*, Nature, Macmillan Publishers Limited, 2013, 18, 528–539.

[3] https://www.hopkinsmedicine.org/ [Last visited 5/11/2017].

[4] Mistry N, Abdel-Fahim R, Samaraweera A, Mougin O, Tallantyre E, Tench C, Jaspan T, Morris P, Morgan P.S, Evangelou N, Imaging central veins in brain lesions with 3-T T2-weighted magnetic resonance imaging differentiates multiple sclerosis from micro angiopathic brain lesions, *Multiple Sclerosis Journal*, 2016, 22, 1289-1296.

[5] Nikas JB, Keene CD, Low WC, Comparison of analytical mathematical approaches for identifying key nuclear magnetic resonance spectroscopy biomarkers in the diagnosis and assessment of clinical change of diseases, *The Journal of Comparative Neurology*, 2010, 518, 4091-4112.

[6] Liu T, Gagan, Agrawal Stratified K-means Clustering Over A Deep Web Data Source, 12th Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, Beijing, China, 2012, 1113-1121.

[7] David SB, Pál D, Simon HU, Stability of k-means clustering, In Proceedings of the 20th annual conference on Learning theory, COLT'07, Berlin, 2007, 20–34.

[8] Xu Y, Qu W, Li Z, Min G, Li K, Liu Z, Efficient *k*-means++ Approximation with MapReduce, *IEEE Transactions on parallel and distributed systems*, 2013, 10,1-10.

[9] Mao Y, Xu Z, Ping P, Wang L, An Optimal Distributed K-Means Clustering Algorithm Based on CloudStack, *Proceedings 2015 Ninth International Conference on Frontier of Computer Science and Technology, China*, 2015.

[10] Yu TK, Lee DT, Chang SM, Zhan J, Multi-party k-Means Clustering with Privacy Consideration, *International Symposium on Parallel and Distributed Processing with Applications*, *Taiwan*, 2010.

[11] Zhao W, Ma H, He Q, Parallel K-Means Clustering Based on MapReduce, *in 09th Proceedings of the 1st International Conference on Cloud Computing*, Springer Beijing, China, 2009, 674-679.

[12] Daniel SM, Anthony FF, John GC, John GM, Randy BL, Open Access Series of Imaging Studies (OASIS): Longitudinal MRI Data in No demented and Demented Older Adults, *Journal of Cognitive Neuroscience*, *MIT Press*, 2010, 22, 2677-2684.

[13] http://www.oasis-brains.org/ [Last visited 5/11/2017].

[14] Elizabeth MS, Russell TS ,Shiee N, Farrah JM, Avni AC, Jennifer LC, Peter AC, Dzung LP, Daniel SR, Ciprian MC, OASIS is Automated Statistical Inference for Segmentation, with applications to multiple sclerosis lesion segmentation in MRI, *NeuroImage: Clinical, Elsevoir*, 2013, 15, 402–413.

[15] Daniel SM, Tracy HW, Parker J, John GC, Randy LB, Open Access Series of Imaging Studies (OASIS): Cross-sectional MRI Data in Young, Middle Aged, Nodemented and Demented Older Adults, *Massachusetts Institute of Technology Journal of Cognitive Neuroscience*, 2007, 19, 1498–1507.

[16] Garner R, WEKA: The Waikato Environment for Knowledge Analysis, Hamilton, Proceedings of the 1995 New Zealand Computer Science Research Students Conference, *Stephen Department of Computer Science, University of Waikato*, 1995, 57–64.

[17] Dean J, Ghemawat S, MapReduce: Simplified Data Processing on Large Clusters, *Communications of the ACM - 50th anniversary*, 2008, 51, 107-113.\

[18] Dittrich J, Arnulfo J, Ruiz Q, Efficient Big Data Processing in Hadoop MapReduce, *Proceedings of the VLDB Endowment VLDB Endowment*, 2012, 5, 2014-2021.

[19] Oancea B, Dragoescu RM, Romama R, Integrating R and Hadoop for Big Data Analysis, *Statistica*,

2014, 2, 83-94.
[20] Kumar S, Singh P, Rani S, Sentimental analysis of social media using R language and Hadoop, 5th *International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, 2016, 7, 207-213.

Authors' Profiles

Tawseef Ayoub Shaikh, research scholar in department of computer engineering, Zakir Hussain College of Engineering & Technology (ZHCET), Aligarh Muslim University Aligarh India. His interests are AI, Machine Learning, Data Analytics and is presently working in Biomedical data analytics field.

Umar Badr Shafeeque, research scholar in department of computer engineering, Zakir Hussain College of Engineering & Technology (ZHCET), Aligarh Muslim University Aligarh India. His interests are Cryptography, Security from the Perspective of Machine Learning Approaches, AI and is presently working in Information Security and Cryptography field.

Maksud Ahamad, research scholar in department of computer engineering, Zakir Hussain College of Engineering & Technology (ZHCET), Aligarh Muslim University Aligarh India. His interests are Soft Computing based machine learning algorithms, AI, Data Mining and is presently working in Soft Computing field.

How to cite this paper: Tawseef Ayoub Shaikh, Umar Badr Shafeeque, Maksud Ahamad,"An Intelligent Distributed K-means Algorithm over Cloudera /Hadoop", International Journal of Education and Management Engineering(IJEME), Vol.8, No.4, pp.61-70, 2018.DOI: 10.5815/ijeme.2018.04.06