International Journal of Engineering & Technology, 7 (4.5) (2018) 685-688

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Analysis of large volume data processing using clustering algorithms

Sarada. B [1] *, Vinayaka Murthy. M [2], Udaya Rani. V [3]

[1] *Research scholar, REVA University, India*
[2] *Professor and Assistant director R&D, REVA University, India*
[3] *Associate Professor, School of C & IT,REVA University, India*
**Corresponding author E-mail: saradasaikonda@gmail.com*

Abstract

The study of large dataset with velocity, variety and volume which is also known as Big data. When the dataset has limited number of clusters, low dimensions and small number of data points the existing traditional clustering algorithms can be used.. As we know this is the internet age, the data is growing very fast and existing clustering algorithms are not giving the acceptable results in terms of time complexity and spatial complexity. So there is a need to develop a new approach of applying clustering of large volume of data processing with low time and spatial complexity through MapReduce and Hadoop frame work applying to different clustering algorithms, k-means, Canopy clustering and proposed algorithm .The analysis shows that the large volume of data processing will take low time and spatial complexity when compared to small volume of data.

Keywords: *Big Data; Canopy Clustering; Hadoop; K-Mean Clustering; Data Processing Techniques; Mapreduce.*

1. Introduction

The data is increasing in terms of volume, variety, and velocity, the existing clustering algorithm takes more time to produce the results. To produce results in terms of less time and less memory one should think of something big and that is parallel programing. MapReduce is one of the programming designs for large volumes of datasets in parallel .MapReduce with HDFS can be used to handle the big data ,which is commonly known as Hadoop .Once the file is placed into HDFS it can be read n number of times.

1.1. Map reduce

MapReduce is a frame work and it is patented by Google which supports processing of large data sets in parallel across Hadoop clusters .The MapReduce is a program block which divides the data and merges the intermediate results. Implementation of MapReduce can be done using any language to run the job [3] .It has two phases namely map and reduce/map () and reduce ().

Map phase

The input applied in this phase is divided into chunks. By default, splitting is done by Hadoop Distributed System (HDFS).The size of the chunks are mutable. The input is key –value in the form of records in MapReduce [3].

The map function takes key and values as input and produces the intermediate values of list as:

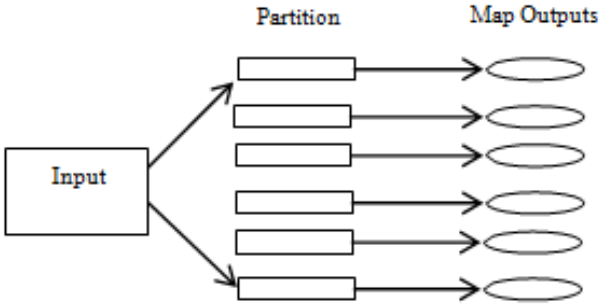

Fig. 1: Map Function Applied to Input.

Map (input key, input value) (output key, output intermediate list) Different data sets produce different intermediate list since the map function runs in parallel.

1.2. Reduce phase

After the completion of map function the intermediate values are combined to get the final result for the same output key. Like map function even reduce function runs in parallel and each of the reduce function run on a different output key [3].

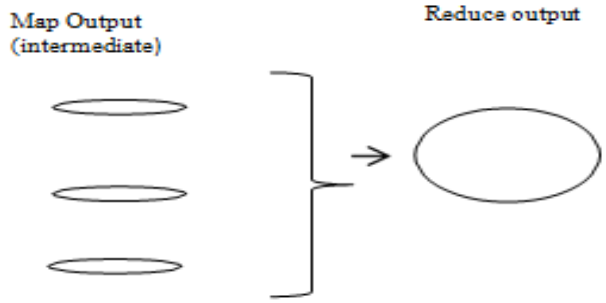

Fig. 2: Reduce Function.

In the above Fig.2 reduce function is applied on map function output which is intermediate list values to get final result

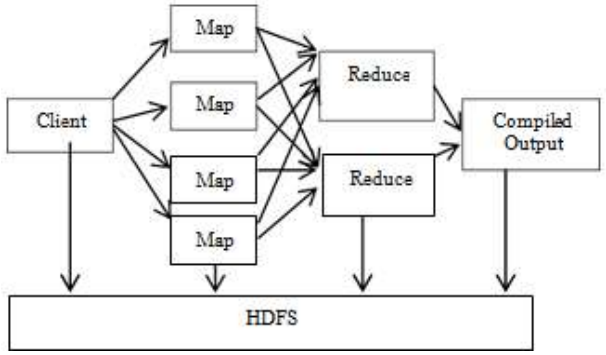

Fig. 3: Map Reduce.

The above Fig.3 shows the combination of Map and Reduce

2. Existing system approach

Clustering is the best example for unsupervised learning algorithm. It is a simple approach to group data points or objects. Here the groups are called clusters. The objects are data points which are in the cluster are similar than those in the other clusters.

2.1. k-means clustering algorithm

K-means clustering algorithm is very simple and easy to understand. The steps involved in this algorithm are:
Step 1: Randomly select the centroids and place them in space, which are temporary means of the cluster.
Step 2: Calculate the Euclidean distance between each data point and cluster center. And then assign the data points to cluster centroid whose distance is minimum.
Step 3: Recalculate the centroids for each cluster and replace by respective cluster centroid.
Step 4: If there is no reassignment of the data point then go to next step otherwise go to step2
Step 5: End

2.2. Limitations

Some of the drawbacks of existing k-mean algorithm through literature survey are:

1) A review of uncertainty handling formalisms by A. Hunter and S. Parsons [6].In this paper computation time is reduced but initial centroids are selected randomly.
2) An overview from a database perspective by M. S. Chen, J. Han, and P. S. Yu. [4]. In this paper author proposed the initial centroid algorithm to avoid selection of random centroid
3) Efficient k-mean clustering algorithm for reducing the time complexity by D.Napoleon, P.Ganga Lakshmi. The authors say that reducing the time complexity is expensive for high dimensional datasets [3].
4) Overcoming the Defects of k-Means Clustering by using Cano- py Clustering Algorithm by Ambika .s and Kavitha G [1]. Avoided random selection of centroid by using canopy clustering algorithm.

3. Proposed system

The main aim of the proposed System is to find the initial values of centroids that is K value for K-means clustering algorithm and studying the space complexity and time complexity on Hadoop and MapReduce platform.
The Modules used in proposed system are

1) Big Data
2) Canopy clustering Algorithm
3) k-Mean Clustering Algorithm

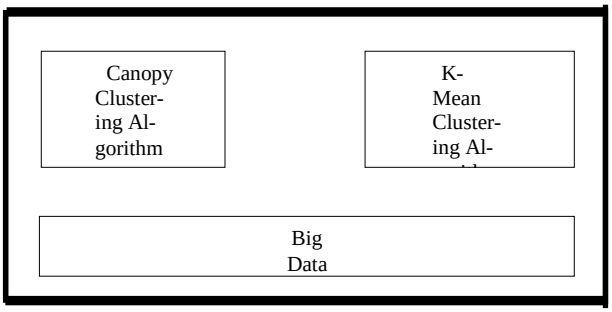

Fig. 4: Proposed System.

3.1. Big data

Big data' is the term used to describe collection of data that is huge in size and yet growing exponentially with time and have the dimensions velocity, variety, volume.

3.2. Canopy clustering algorithm

The results of this algorithm are a number of canopies which are the cluster centers for the given dataset.

3.3. K-mean clustering algorithm

The execution time of K-Mean clustering Algorithm Given by O (nkdi) where n is the number of data points, k is the number of clusters, i is the number of iterations needed to converge and d is the dimensions. When the value of n and d increases then it is time consuming process or it is not applicable .In order to overcome the canopy clustering algorithm is used which is also called as pre clustering algorithm. In the Proposed system the output of the canopy clustering algorithm is given as input to the k-mean clustering algorithm

3.4. Canopykmeans clustering algorithm

Input: Dataset
Output: number of clusters
The algorithm uses two threshold values T1 and T2 Where T1 is loose distance and T2 is tight distance Where T1>T2
The steps involved in canopy clustering algorithm are:
Step 1: Randomly select any data point from the dataset as a canopy center
Step 2: Find the distance to all other points in the dataset from the canopy center.
Step 3: The distance calculated is less than the T1 then put data points into a canopy
Step 4: Remove from the data set all the points which are less than T2
Step 5: Repeat the above step1 to step 4 until the dataset becomes empty
Step 6: Feed the output as input K-mean clustering algorithm

3.5. Result and analysis

Table 1: Time and Spatial Complexity of K-Means ALG

K Means Clusering		
Data	Timing Complexity	Spatial Complexity
5000	20	0.05
10000	21	0.047619048
15000	21	0.047619048
20000	21	0.047619048
25000	21	0.047619048
30000	21	0.047619048
35000	21	0.047619048
40000	21	0.047619048
45000	21	0.047619048
50000	21	0.047619048
55000	21	0.047619048
60000	21	0.047619048
65000	21	0.047619048
70000	21	0.047619048
75000	28	0.035714286
80000	28	0.035714286
85000	28	0.035714286
90000	28	0.035714286
95000	28	0.035714286
100000	28	0.035714286

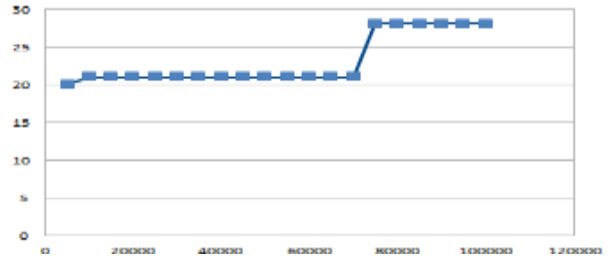

Fig. 5: K Means Clustering Time Complexity.

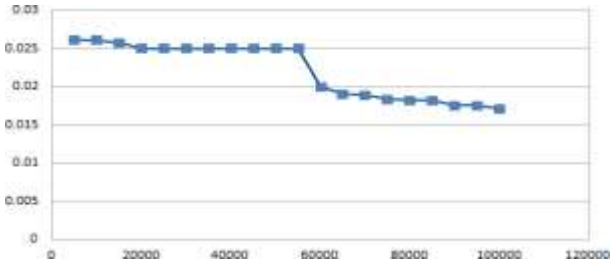

Fig. 6: K Means Clustering Spatial Complexity.

Table 2: Time and Spatial Complexity of Canopy ALG

Canopy Algorithm		
Data	Timing Complexity	Spatial Complexity
5000	38.25	0.026143791
10000	38.36	0.026068822
15000	38.9	0.025706941
20000	40	0.025
25000	40	0.025
30000	40	0.025
35000	40	0.025
40000	40	0.025
45000	40	0.025
50000	40	0.025
55000	40	0.025
60000	50	0.02
65000	52.5	0.019047619
70000	53	0.018867925
75000	54.45	0.018365473
80000	55	0.018181818
85000	55	0.018181818
90000	57	0.01754386
95000	57	0.01754386
100000	58.5	0.017094017

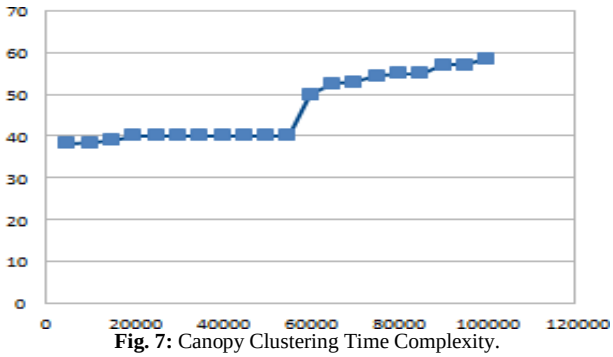

Fig. 7: Canopy Clustering Time Complexity.

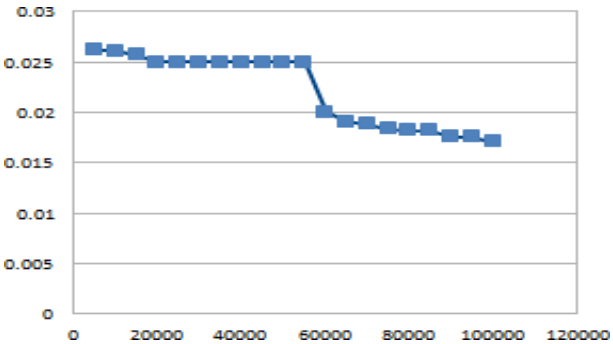

Fig. 8: Canopy Clustering Time Complexity.

Table 3: Time and Spatial Complexity of Proposed System

OVERALL OUTPUT		
Data	Timing Complexity	Spatial Complexity
5000	50	0.02
10000	60	0.016666667
15000	60	0.016666667
20000	63	0.015873016
25000	65	0.015384615
30000	75	0.013333333
35000	80	0.0125
40000	80	0.0125
45000	80	0.0125
50000	80	0.0125
55000	82	0.012195122
60000	82	0.012195122
65000	86.5	0.011560694
70000	87.25	0.011461318
75000	88.5	0.011299435
80000	89.5	0.011173184
85000	89.5	0.011173184
90000	89.5	0.011173184
95000	89.5	0.011173184
100000	89.5	0.011173184

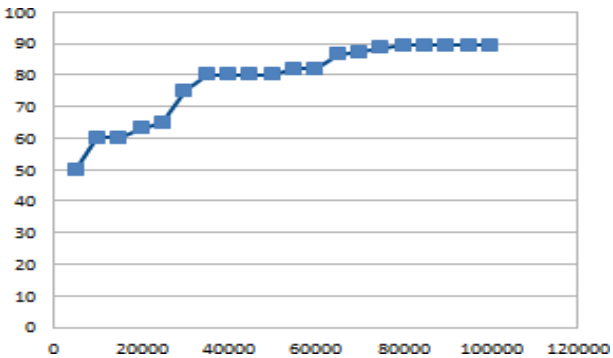

Fig. 9: Proposed System Time Complexity.

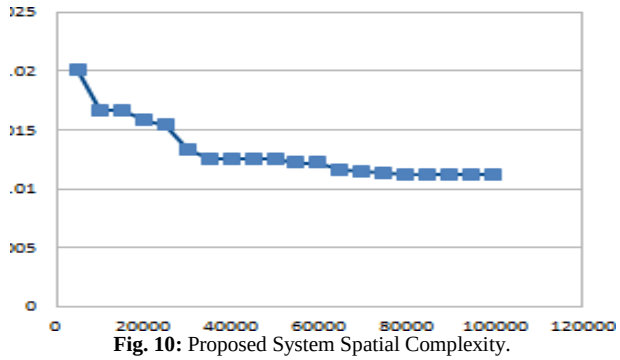

Fig. 10: Proposed System Spatial Complexity.

The data is growing in terms of volume, variety and velocity .The behavior of each clustering algorithm is analyzed through MapReduce and Hadoop platform which uses parallel processing technique. Here we considered the simulated social data of size one lakh with twelve attributes. The Figures 5 ,6 and Table1 shows that as the dataset increases the time taken and spatial com- plexity for k-mean clustering algorithm is less and constant as the is increasing and the same result from Figure 7,8 and Table 2 that is canopy clustering algorithm. The Proposed approach takes
spatial complexity less than the canopy and k-mean clustering algorithm from the Figure 9, 10 and Table 3 and no need to give K value manually.

4. Conclusion

In this paper we have studied existing K-mean and canopy clustering algorithms for big data using MapReduce and Hadoop platform. And proposed new technique, the canopy algorithm is applied to the Big data and the output is given as the initial centers (the value of k)to K-mean clustering algorithm through MapReduce and Hadoop frame work which uses parallel processing technique.

Acknowledgement

I would like to express my special thanks of gratitude to my guide Dr.M.Vinayaka Murthy and Dr.Udaya Rani.V for their continuous support to do research paper and also my husband Saikonda Venkateswarlu who has helped with technical guidance to do the research paper.

References

[1] Ambika.s and Kavitha.G," Overcoming the Defects of K-means clustering by using Canopy Clustering Algorithm IJSRD |*Vol. 4, Issue 05, 2016 | ISSN (online): 2321-0613.*
[2] D. Napoleon & P. Ganga lakshmi "An Efficient K-Means Clustering Algorithm for Reducing Time Complexity Using Uniform Distribution Data Points" IEEE, 2010, pp, 42-45.
[3] Dweepna Garg 1, Khushboo Trivedi 2, B.B.Panchal ," A Comparative study of Clustering Algorithms using MapReduce *2321-0613* in Hadoop" *IJSRDt| Vol. 4, Issue 05,2016 | ISSN (online):*
[4] M. S. Chen, J. Han, and P. S. Yu. IEEE Trans Knowledge and Data Engineering Data mining. An overview from a database perspective, 8:866-883, 1996.
[5] Ayman E. Kheer, Ahmed I. El Seddawy, Amira M. Idrees," Performance Tuning of K-Mean Clustering Algorithm a Step towards Efficient DSS", IJIRCST,ISSN: 2347-5552, Volume 2, Issue 6, November – 2014.
[6] A. Hunter and S. Parsons, "A review of uncertainty handling formalisms", Applications of Uncertainty Formalisms LNAI 1455, pp.8-37. Springer – Verlag, 1998.
[7] H.R. Shashidhar, G.T. Raju and M Vinayaka Murthy,"Efficient Estimation of Result Selectivity for Web Query Optimi zation", International Journal of Pure and Applied Mathematics, Volume 17 No. 7 2017, PP 193-205, ISSN:311-8080.
[8] H.R. Shashidhar, G.T. Raju and M Vinayaka Murthy, "Effective Cost Models for Web Query Optimization", International Journal of Pure and.
[9] Applied Mathematics, Volume 117 No. 20, 17, PP 727-739, ISSN: 1311-8080.
[10] M Vinayaka Murthy "Survey On Web Query Optimization Trends and Future Research", International Conference On Advanced Material Technology 2016, Issue –V, pp 409 – 417, Elsevier Materials Today: Proceedings.
[11] M Vinayaka Murthy, "A Comparative Study on Mining the & Healthy Food Preferences of Women Clusters", Journal of Scientific Engineering Research, Vol 6, Issue 7, pp 126 131, 2017, ISSN: 2229-5518.