Scientific Journal of Impact Factor (SJIF): 5.71

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 5, Issue 03, March -2018

Two Level Clustering Using Hadoop Map Reduce Framework

Ankita Dubey, Dr. Abha Choubey

[1] *Shri Shankaracharya Group of Institutions,*
Dept. of Computer Science and Engineering, Bhilai, Chhattisgarh, India

[2]*Guide, Associate Professor, Shri Shankaracharya Group of Institutions,*
Dept. of Computer Science and Engineering, Bhilai, Chhattisgarh, India

Abstract— *In the field of information mining, clustering is one of the vital techniques. K-Means is an ordinary separation based clustering calculation; 2-level clustering should actualize adaptable clustering by methods for partitioning, inspecting and information coordinating. Among those apparatuses of disseminated handling, Map-Reduce has been generally grasped by both scholarly world and industry. Hadoop is an open-source parallel and conveyed programming system for the usage of Map-Reduce figuring model. With the investigation of the Map-Reduce worldview of figuring, we find that Hadoop parallel and disseminated registering model is proper for the usage of adaptable clustering calculation. This paper takes focal points of K-Means, 2-level clustering component and Map-Reduce registering model; proposes another technique for parallel and circulated clustering to investigate dispersed clustering issue in view of Map-Reduce. The strategy intends to apply the clustering calculation successfully to the disseminated condition. The broad investigations show that the proposed calculation is adaptable, and the time execution is steady.*

Keywords— *Hadoop, MapReduce, K-Means Clustering, Two Level.*

I. INTRODUCTION

Clustering is one of the essential strategies in the field of information mining; clustering is a strategy for unsupervised learning, and is connected to information investigation in many fields, including machine learning, design acknowledgment, picture examination and insights [1]. Bunch investigation is to dole out an arrangement of items into subsets (called groups) with the goal that articles in a similar bunch have higher comparability and protests in the distinctive bunches have bring down similitude. The objective of clustering is to group the information in light of closeness.

K-Means is a clustering calculation which has been known as a point of reference, it might be by a wide margin the most generally utilized information mining instrument in commonsense applications. K-Means calculation was positioned second of best 10 calculations in information mining by the ICDM Conference in October 2006, which C4.5 was positioned first [2]. Notwithstanding, with the improvement of the Internet, and in addition the thought for security insurance, the circulated stockpiling qualities of information are progressively huge, and various clustering calculations cannot be all around connected to the appropriated condition. Consequently, it is of fundamental significance and pressing need to consider how we can apply K-Means calculation to the conveyed condition.

Enormous datasets are getting to be plainly pervasive [3]; as the extent of accessible datasets has developed from Megabytes to Gigabytes and now into Terabytes, machine learning calculations and registering frameworks have been ceaselessly advancing with an end goal to keep pace. Fascinating true applications deliver colossal volumes of information, thusly how to apply dispersed registering model on a critical mining undertaking has been a prominent issue. As information turn out to be more rich and versatile, devices for dispersed preparing are likewise rising. As a matter of fact, there are a considerable measure of procedures for appropriated figuring, for example, groups, network and P2P strategies. Dispersed figuring mode can be a leap forward contrasted with desktop registering mode and multi-center innovation [4]. Thusly, on the off chance that we send the assignment of information mining to an open circulated framework, it is a want to accomplish versatile information mining totally.

A. K-Means Clustering

Clustering is important and essential concept of data mining field used in various applications. In Clustering, data are divided onto various classes. These classes represents some important features. Means, classes are the container of similar behavior of objects.

The objects which behave or are closer to each other are grouped in one class and who are far or non-similar are grouped in different class. Clustering is a process of unsupervised learning. Highly superior clusters have high intra-class similarity and low inter-class similarity.

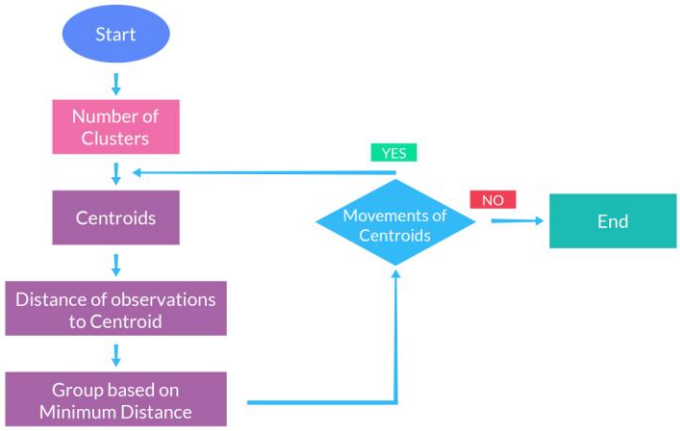

Fig.1. K-Means Generic Algorithm

II. MAP REDUCE PROGRAMMING MODEL

Map-Reduce programming model is composed by Dean et al. [5], which is utilized as a part of parallel and dispersed handling condition to manage tremendous measures of information estimations. This model partitions an assignment into more duplicates of fine-grain subtasks. After these sub-assignments are dispatched and handled among the sit preparing hubs, it creates the last outcomes through particular tenets to consolidate. Map-Reduce preparing model is to some degree like deterioration and acceptance of conventional programming model. It is anything but difficult to utilize, and gives stack adjusting and blame tolerant instrument.

The model edited compositions disseminated processing into two stages of Map and Reduce keeping in mind the end goal to accomplish proficient dispersed applications. The projects composed with Map-Reduce have dispersed properties, which can make parallel and conveyed handling on groups, framework and distributed computing situations. What Developers need to do is to accomplish its own Map and Reduce capacities, and afterward to submit to the Map-Reduce working condition. For instance, Hadoop is an execution of the open-source structure for Map-Reduce parallel figuring model.

The 2-level clustering component is proper for parallel and appropriated registering, specifically for MapReduce. This paper uses K-Means, 2-level clustering component and Map-Reduce figuring model, and proposes another technique Instances K-Means for parallel and dispersed registering. Fundamental commitments of this paper can be outlined as takes after.

(1) To actualize the appropriated processing model, we propose 2-level clustering component what's more, another strategy Instances K-Means in view of K-Means calculation;
(2) We utilize the Map-Reduce figuring model to apply the clustering calculation to the circulated condition and assemble a group with 1 machines keeping in mind the end goal to perform tests;
(3) Our test comes about demonstrate that our way to deal with accomplish adaptable information mining is proficient, and including number of hubs would profit the time execution of clustering.

III. LITERATURE SURVEY

K. A. Abdul Nazeer et al. [6] proposes k-means algorithm, for different sets of values of initial centroids, produces different clusters. Final cluster quality in algorithm depends on the selection of initial centroids. Two phases includes in original k means algorithm: first for determining initial centroids and second for assigning data points to the nearest clusters and then recalculating the clustering mean.

Soumi Ghosh et al. [7] present a comparative discussion of two clustering algorithms namely centroid based K-Means and representative object based FCM (Fuzzy C-Means) clustering algorithms. This discussion is on the basis of performance evaluation of the efficiency of clustering output by applying these algorithms.

Shafeeq et al. [8] present a modified K-means algorithm to improve the cluster quality and to fix the optimal number of cluster. As input number of clusters (K) given to the K-means algorithm by the user. But in the practical scenario, it is very difficult to fix the number of clusters in advance. The method proposed in this paper works for both the cases i.e. for known number of clusters in advance as well as unknown number of clusters. The user has the flexibility either to fix the number of clusters or input the minimum number of clusters required.

Junatao Wang et al. [9] propose an improved k-means algorithm using noise data filter in this paper. The shortcomings of the traditional k-means clustering algorithm are overcome by this proposed algorithm. The algorithm develops density based detection methods based on characteristics of noise data where the discovery and processing steps of the noise data are added to the original algorithm. By pre-processing the data to exclude these noise data before clustering data sets the cluster cohesion of the clustering results is improved significantly and the impact of noise data on k-means algorithm is decreased effectively and the clustering results are more accurate.

Shi Na et al. [10] present the analysis of shortcomings of the standard k-means algorithm. As k-means algorithm has to calculate the distance between each data object and all cluster centers in each iteration. This repetitive process effects the efficiency of clustering algorithm. An improved k-means algorithm is proposed in this paper.

IV. METHODOLOGY

In this section the proposed system architecture with detailed explanation are discussed. Fig. 2. Shows the proposed system architecture.

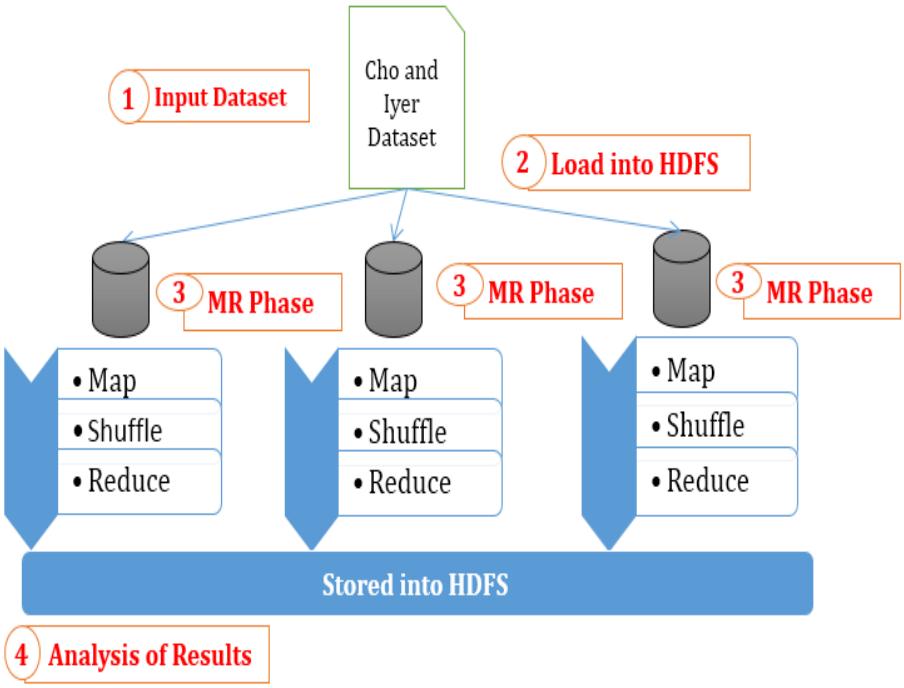

Fig. 2. Proposed system work flow

There are various modules which are responsible for clustering documents. The k-means clustering algorithm is presented in map and reduce phase in fig. 3 and 4.

A. Cho and Iyer Dataset

Cho and Iyer dataset are gene sequences dataset which need to be clustered according to the similar gene patterns. More about dataset are described in result section.

B. HDFS

Both the dataset should be loaded into hadoop distributed file system for processing. The HDFS breaks the dataset into smaller pieces for parallel operation.

C. MR Phase

The algorithm of parallel k-means are implemented in Map and Reduce phases of hadoop. The algorithm is presented below.

D. Map Phase

Fig.3 shows the process of a Mapper. This class implements Map operation in Mapper interface, and the input parameter value is instance type. Input Reader divides the input file into m splits. Each Mapper worker is applied to each split from Input Reader according to the format of the (key1, value1) pairs, each call produces a list (key2, value2) of (key2, value2) pairs as the intermediate results.

Each Mapper operation is responsible for generating the first clustering results, key1 is the file name corresponding to the input dataset, and value1 is a formatted text of Instances type through the division of the input file. According to the K-Means algorithm, Map function builds first clustering, and then Emit function submits the intermediate results list (key2, value2) to Partition function. Since then, the tasks of Map phase are completed.

function mapper(key_1, $value_1$)

1: Read one of *m* splits from Input Reader;
/* *Build first clustering* */
2: Place *k* points represented initial group centroids;
3: Assign each object $value_1$ to the group with the nearest centroid;
4: Recalculate the positions of the *k* centroids;
5: Repeat Step 3 and 4 until the centroids no longer move;
6: $value_2$ = all centroids;
/* *Submit intermediate results* */
7: ***Emit*** *list* (key_2, $value_2$) **to** Partition function;

Fig. 3. Shows the Mapper Working

E. Reduce Phase

Fig.4. shows the process of a Reducer. This class implements Reduce operation in Reducer interface. Partition function collects all the (key2, value2) pairs with the same key in the list and integrates them together to produce (key2, list (value2)). list (value2) is a set of intermediate results, which is submitted by Partition function.

The intermediate results produced by Map function are partitioned into r regions and assigned to Reducer workers, each Reducer worker is applied in parallel to each split list (value2) based on K-Means algorithm. Reduce operation makes (key2, list (value2)) as inputs, each call of Reducer worker generates one value3 or an empty return according to the idea of 2-tier clustering. The results of all calls form the result list of list (value3).

function reducer(key_2, *list* ($value_2$))

1: Read one of *list* ($value_2$) from Partition function;
/* *Build second clustering* */
2: Place *k'* (*k'*<*k*) points represented initial group centroids;
3: Assign each object $value_2$ to the group with the nearest centroid;
4: Recalculate the positions of the *k'* centroids;
5: Repeat Step 3 and 4 until the centroids no longer move;
6: Produce $value_3$ according to integration strategy of 2-tier clustering;
/* *Submit final results* */
7: ***Emit*** *list* ($value_3$) **to** Output Writer;

Fig. 4. Shows the Reducer Working

V. RESULT

In this section, result of proposed algorithm and simulation are presented. Basically we worked on two tier clustering algorithm named parallel k-means clustering.

Cho and Iyer dataset are collected from UCI machine learning repository. Dataset snapshot is attached in fig. 5. The format of the dataset are:

Each row represents a gene:

1) The first column is gene_id.
2) The second column is the ground truth clusters. You can compare it with your results. "-1" means outliers.
3) The rest columns represent gene's expression values (attributes).

```
2   1   -0.21   0.19    0.86    0.04    -0.35   -0.39
-1.23   -0.325  0.0
3   1   -0.3    -0.56   -0.29   -0.5    -0.27   -0.29
-0.12   -0.16   0.67
4   1   0.07    0.26    -0.47   -0.68   -0.63   -0.39
-0.2    -0.06   0.36
5   1   -1.04   0.13    0.51    -0.44   -0.88   -0.32
-0.13   0.092   0.0
6   1   -1.17   0.09    -0.52   -1.04   -1.16   -0.83
0.2 0.91    0.68
7   1   -0.16   0.35    -0.13   -0.26   -0.4    -0.47
-0.59   0.14    0.2
```

Fig. 5. Snapshot of cho.txt dataset

After applying clustering algorithm, the p-k-means clustering algorithm performs as shown in table I.

TABLE I. shows the k-means clustering input and output data

Characteristics	**CHO**	**IYER**
Dataset Size	386 Gene Profile	517 Gene Profile
Centroid Points	5 Data points	10 Data points
Convergence	11th iteration	11th iteration
Clustered Points	5	10

Fig. 6. And 7. Shows the output of cho and iyer clustered output data points.

```
1  0   -1.16   -1.39   -0.96
   -0.91   -0.71   -0.1    0.54
   0.61    0.63    0.25    0.37
   -0.33   0.1 0.493   1.27
   0.87    -0.3    -0.56   -0.29
   -0.5    -0.27   -0.29   -0.56
   -1.04   0.32    0.9 0.45
   0.17    0.164   -0.12   -0.16
   0.67    0.07    0.26    -0.47
   -0.68   -0.63   -0.39   0.07
   0.79    0.58    0.31    -0.14
   -0.29   -0.103  -0.2    -0.06
   0.36    -1.04   0.13    0.51
   -0.44   -0.88   -0.32   0.21
   0.95    1.07    0.38    0.01
   -0.13   -0.78   -0.13   0.092
```

Fig. 6. Cluster 1 output of Cho dataset

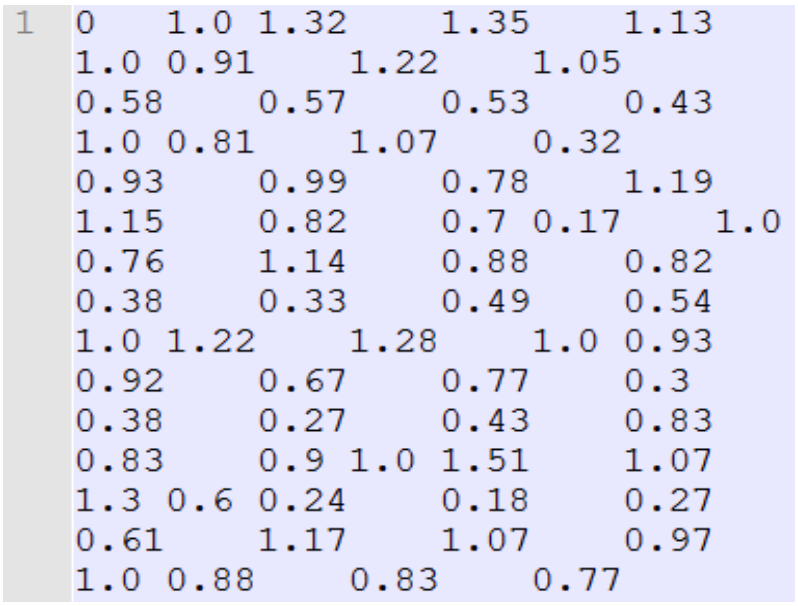

Fig. 7. Cluster 1 output of Iyer dataset

VI. Conclusion

In this paper, we present a novel mechanism in which we have utilized hadoop map reduce to operate in parallel. The k-means algorithm is implemented in map reduce phase to make it more efficient while processing. The proposed method effectively utilizes the parallel approach of map reduce and produces results in shorter duration of time.

VII. References

[1] U. Fayyad, R. Uthurusamy, "Data Mining and knowledge discovery in databases," Communications of the ACM, vol. 39, no. 11, pp. 24-26, November 1996.
[2] X. Wu, V. Kumar, Ross, J. Ghosh, Q. Yang, H. Motoda, G. Mclachlan, A. Ng, B. Liu, P. Yu, Z.-H. Zhou, M. Steinbach, D. Hand, and D. Steinberg, "Top 10 algorithms in data mining," Knowledge and Information Systems, vol. 14, no. 1, pp. 1-37, January 2008.
[3] S. Papadimitriou, J. Sun, "DisCo: Distributed Co-clustering with Map-Reduce: A Case Study Towards Petabyte-Scale End-to-End Mining," in: Eighth IEEE International Conference on Data Ming, pp. 512-521, December 2008.
[4] K. Cardona, J. Secretan, M. Georgiopoulos, G. Anagnostopoulos, "A Grid Based System for Data Mining Using MapReduce," Technical Report, University of Puerto Rico, July 2007.
[5] J. Dean, S. Ghemawat, "MapReduce: Simplified data processing on large clusters," in: Proceedings of the 6th Symposium on Operating System Design and Implementation. San Francisco, California, USA. USENIX Association, pp. 137-150, December 2004.
[6] K. A. Abdul Nazeer, M. P. Sebastian,îImproving the Accuracy and Efficiency of thek-means Clustering Algorithm, Proceedings of the World Congress on Engineering 2009 Vol I WCE 2009, July 1 - 3, 2009, London, U.K.
[7] Soumi Ghosh, Sanjay Kumar Dubey, Comparative Analysis of K-Means and Fuzzy C-Means Algorithmsî, International Journal of Advanced Computer Science and Applications, Vol. 4, No.4, 2013
[8] Shafeeq,A., Hareesha,K.,ìDynamic Clustering of Data with Modified K-Means Algorithm, International Conference on Information and Computer Networks, vol. 27 ,2012
[9] Junatao Wang, XiaolongSu,îAn Improved K-means Clustering Algorithm, Communication Software and Networks (ICCSN), 2011 IEEE 3rd International Conference on 27 may,2011 (pp. 44-46)
[10] Shi Na, Liu Xumin, Guan Yong, ìResearch on K-means Clustering Algorithm: An Improved K-means Clustering Algorithm, Intelligent Information Technology and Security Informatics,2010 IEEE Third International Symposium on 2-4 April, 2010(pp. 63-67)

