ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology
Vol. 2, Issue 8, August 2015

Data Clustering using MapReduce for Multidimensional Datasets

N.Vishnupriya[1], Dr.F.Sagayaraj Francis[2]

PG Student, Department of Computer Science and Engineering, Pondicherry Engineering College, Pondicherry [1]

Professor, Department of Computer Science and Engineering, Pondicherry Engineering College, Pondicherry[2]

Abstract: Data mining techniques make possible to analyze and discover knowledge of data sets. However, the tradition clustered data are not providing more accurate data for large datasets. Mahout support for implementing cluster algorithms by handling large volume of data in integration with hadoop. Using MapReduce programming model for processing the data cluster in distributed systems. To improve the performance of the large-scale datasets clustering on the single computer. To find the accuracy of data in K- Mean's algorithm to calculate SSE value based upon Euclidean distance using MapReduce framework for 2dimension and 3 dimension datasets.

Keywords: Clustering, Hadoop, K-means, MapReduce, SSE (sum of square error).

I. INTRODUCTION

Data clustering is the partitioning of object into groups (called clusters) such that the similarity between objects of the same group is maximized and similarity between objects of different groups is minimized. The goal of the clustering technique is to decompose or partition a data set into groups such that both intragroup similarity and inter-group dissimilarity is maximized. Each clustering algorithm is based on some kind of distance measures, which leads to grouping of related objects. The distance measure is used to determine similarity of object criteria. As each distance measure shows different methods for defining the degree of comparison between two objects.The K-Means algorithm uses Euclidean distance to measure the distortion between a data object and its cluster centroid. Euclidean distance metric is sufficient to successfully group similar data instances. K-Means clustering is a method used to of the most commonly and effective methods to classify data because of its simplicity and ability to handle voluminous data sets.

A. Hadoop

Hadoop Distributed File System (HDFS) is the storage component of the Hadoop framework that has been designed for processing and maintaining huge datasets efficiently among cluster nodes. Hadoop became the fastest system to sort a terabyte of data. A small Hadoop cluster includes a single master and multiple worker nodes. The master node consists of a Task Tracker, JobTracker, DataNode and NameNode. A slave or worker node acts as both a DataNode and TaskTracker, though it is possible to have data-only worker nodes and compute-only worker nodes.

B.MapReduce Programming

In MapReduce process has two separate steps Map and Reduce steps. Each step is process on sets of (key, value) pairs. While, the time of program execution is divided into a Map and a Reduce stage, each separated by data transfer between nodes in the cluster. In Mapper function can select the data values as input, applies the function to each value to the given datasets and generates an output set. The mapper output in the form of (key, value) pairs. The framework, then, sorts the mapper function outputs and inputs them into a Reducer. This data transfer between the Mappers and the Reducer. The values are combined at the node running the Reducer for that key. In Reducer stage produces another set of (key, value) pairs as final output. The Reducer stage can only process after all data get from the Map process. MapReduce requires the input as a (key, value) pair that can be divided and therefore, limited to tasks and algorithms that use (key, value) pairs.

II.RELATED WORK

Fahim.A.M et al [3], proposed an idea that makes k-means more efficient, especially for dataset containing large number of clusters. In each iteration, the k-means algorithm computes the distances between data point and all centers, this is computationally very expensive for huge datasets. For each data point, we can keep the distance to the nearest cluster. At the next iteration, we compute the distance to the previous nearest cluster. At the next iteration, distance to the previous nearest cluster is computed. If the new distance is less than or equal to the previous distance, the point stays in its cluster and there is no need to compute its distances to the cluster centers. This saves the time required to compute distances to k-1 cluster centers. Two functions are written in the proposed method. The first function is the basic function of the k-means algorithm, which finds the nearest center for each data point. Second function is distance new. Here an efficient implementation method for implementing the k-means method is proposed. The algorithm produces the same clustering results as that of the k-means algorithm, and has significantly superior performance than the k-means algorithm.

Dean J et al [4], discussed about data processing on large clusters using MapReduce. MapReduce is a programming model and an associated implementation for processing and generating large data sets. Users specify a mapfunction that processes a key/value pair to generate a set of intermediate key/value pairs, and a reduce function

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
Vol. 2, Issue 8, August 2015

that merges all intermediate values associated with the same intermediate key. The run-time system takes care of the details of partitioning the input data, scheduling the program's execution across a set of machines, handling machine failures, and managing the required inter-machine communication. This allows programmers without any experience with parallel and distributed systems to easily utilize the resources of a large distributed system. MapReduce runs on a large cluster of commodity machines and is highly scalable: a MapReduce process computation many terabytes of data on thousands of machines. Programmers find the system easy to use: hundreds of MapReduce programs have been implemented and upwards of one thousand MapReduce jobs are executed on Google's clusters every day.

KennSlagter et al [5], The proposed algorithm improves load balancing as well as reduces the memory requirements. Nodes which runs slow degrade the performance of Mapreduce job. To overcome this problem, a technique called micro partitioning is used that divide the tasks into smaller tasks greater than the number of reducers and are assigned to reducers in "just-in-time" fashion. Running many small tasks lessens the impact of stragglers, since work that is scheduled on slow nodes is small which can be performed by other idle workers.

Juntaowang et al [6], developed 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. Preprocessing the data improves the clustering result significantly and the impact of noise data on k-means algorithm is decreased. First a pretreatment is made with the data to be clustered to remove the outliers using outlier detection method based on LOF, so that the outliers cannot participate in the calculation of the initial cluster centers, and excluded the interference of outliers in the search for the next cluster center point. We secondly apply fast global k-means clustering algorithm on the new data set which is generated previously. Fast global k-means clustering algorithm is an improved global k-means clustering algorithm by Aristides Likas.

Cluster analysis has undergone dynamic development in the recent years. There are several types of clustering hierarchical clustering, density-based clustering, grid based clustering, model-based clustering and partition clustering. Each clustering type has its own style and optimization methods. Regarding the expansion of the data size and the limitation of a single machine, a natural solution is to consider parallelism in a distributed computational environment. MapReduce is a programming framework for processing large-scale datasets by exploiting the parallelism among a cluster of computing nodes. MapReduce gains popularity for its simplicity, flexibility, fault tolerance and scalability. All the big data processing problems cannot be made efficient by parallelism because partition clustering algorithm requires exponentially much iteration. Also job exponential creation time and time of big data shuffling are hard to accept especially when data size is huge, so just parallelism is not enough, only by eliminating the partition clustering algorithms dependence on the iteration, high performance can be achieved.

III.K-MEANS ALGORITHM

Clustering is a process of grouping with similar objects. Any cluster should exhibit two main properties that belong to, low inter-class similarity and high intra-class similarity. Clustering techniques used to group a large number of things together into clusters that share some similarity. It's a method to discover hierarchy and order in a large or hard to understand datasets and in that way reveal are interesting patterns or make the data set easier to comprehend. Cluster analysis is used in many numbers of applications such as image processing and data analysis.

K-Means is one of the unsupervised learning methods among partitions based clustering methods. It classifies a given dataset of n data objects in k clusters, where k is the number of desired clusters. The K-means algorithm gave better results only when the initial partition was close to the final solution.

K-means clustering algorithm follows the blow steps.

i) Choose a number of desired clusters, *k*.
ii) Choose *k* starting points to be used as initial estimates of the cluster centroids. The initial starting values.
iii) Examine each point (i.e., job) in the workload dataset and assign it to the cluster whose centroid is nearest to it.
iv) When each point is assigned to a cluster, recalculate the new k centroids.
v) Repeat steps 3 and 4 until no point changes its cluster assignment, or until a maximum number of passes through the data set is performed.

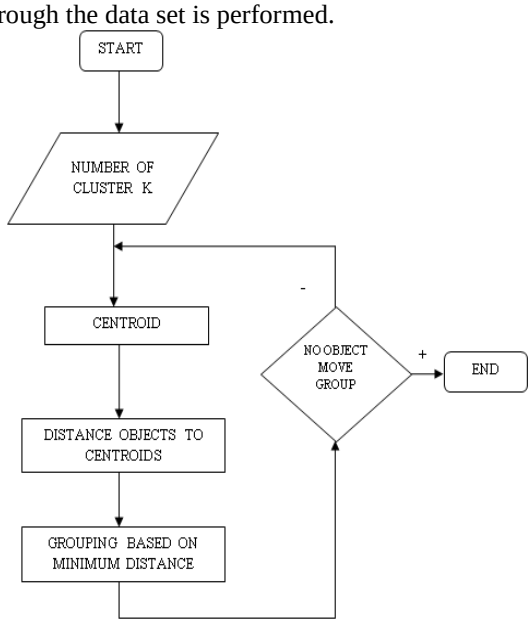

Fig. 1 K-Means Algorithm

IV. PROPOSED METHOD

In proposed method using k-means clustering algorithm to cluster the data for different type of

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
Vol. 2, Issue 8, August 2015

dimensional dataset in Hadoop framework and calculate the SSE value for those data. The k-means algorithm is one of the most effective algorithms for clustering. To find the accuracy to calculate SSE value while calculating the SSE value is small the given dataset is compact. The implementation of clustering algorithm also benefits from the possibility to access by the map reduce framework, so user can use the algorithm with large datasets.

C. Sum of Squared Error (SSE)

The implemented k means clustering algorithm in MapReduce paradigm based upon the Euclidean distance the result of cluster value can calculated by SSE to identify the accuracy of cluster.

$$\sum_{i=1}^{i=n}(((x_i - x_c)\text{^}2 + ((y_i - y_c)\text{^}2)))$$

xi--> x co-ordinate of the points in the cluster.
xc--> x coordinate of the centroid.
yi--> y co-ordinate of the point in the cluster.
yc--> y co-ordinate of the centroid.

D. Architecture Diagram

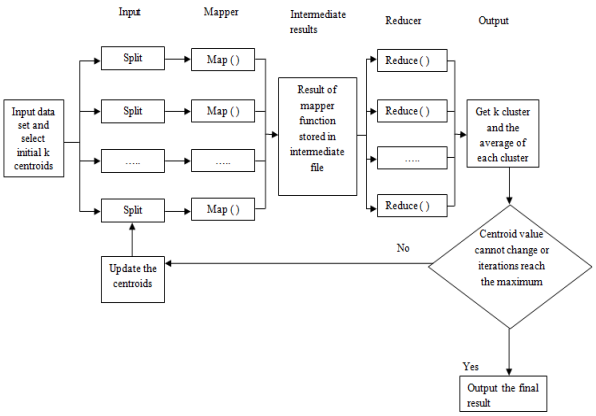

Fig. 2 Process of MapReduce using K-means Algorithm

The K-means clustering algorithm is process by using MapReduce can be divided into the following phases:

1. Initial

(i) The given input data set can be split into sub datasets. The sub datasets are formed into <Key, Value> lists. And these <Key, Value> lists input into map function.

(ii) Select k points randomly from the datasets as initial clustering centroids.

2. Mapper

a) Update the cluster centroids. Calculate the distance between the each point in given datasets and k centroids.

b) Arrange each data to the nearest cluster until all the data have been processed.

c) Output <ai, zj> pair. And ai is the center of the cluster zj.

3. Reducer

(i) Read <ai, zj> from Map stage. Collect all the data records. And then output of k clusters and the data points.

(ii) Calculate the average of each cluster which is selected as the new cluster center.

(iii) Calculate the new centroids with the original centroids in the same cluster. If the value is smaller than the threshold or the number of iterations of the algorithm has reached the maximum, the algorithm will stop. Otherwise, the new cluster centroids points are used to update the original centroids. Return to map stage, and continue the algorithm until merging.

V.EXPERIMENTAL RESULTS

Clustering the data in hadoop framework for k means clustering algorithm based upon the Euclidean distance. Dataset consists of ten thousand rows and two columns in the form of real numbers in structure format. In this paper using different dimensional dataset to calculate the SSE value. Depending upon the centroid the cluster value can be calculated from dataset.

E. Two Dimensional Dataset

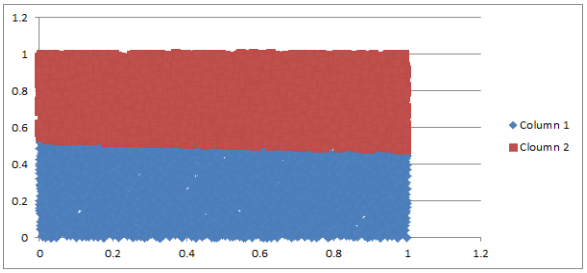

Fig. 3. Dataset for 2Dimensional

TABLE I: SSE VALUE FOR 2DIMENSION DATASET

SSE Value For Two Dimensional									
	K value=2		K value=3		K value=4				
Cluster value	518.241	529.229	113.592	229.460	336.09	195.243	32.382	49.88	284.524
SSE value	1049.466		678.147			562.032			

TABLE I represented the value for 2dimensional dataset and calculate the SSE value for different k value using k means clustering algorithm.

F. Three Dimensional Dataset

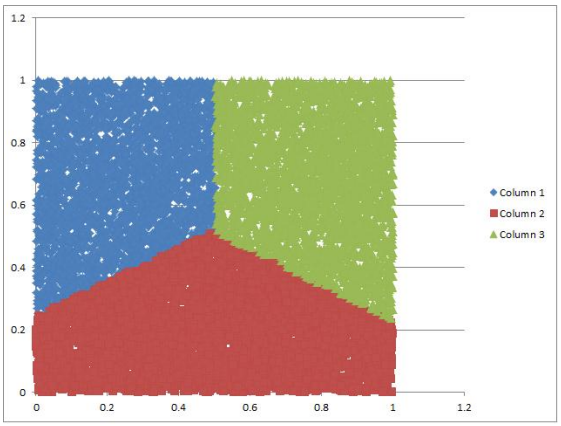

Fig. 4 .Dataset for 2Dimensional

TABLEII represented the value for 3dimensional dataset and calculate the SSE value for different k value using k means clustering algorithm.

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
Vol. 2, Issue 8, August 2015

TABLE II: SSE VALUE FOR 3DIMENSION DATASET

SSE Value For Three Dimensional									
K value=2			K value=3			K value =4			
Cluster value	469.173	479.236	338.541	153.110	273.365	213.008	85.547	130.913	226.459
SSE value	948.409		765.017			655.928			

K-means algorithm which implemented by the Euclidean distance. The SSE value is low the cluster data is highly accuracy means the selection of particular distance measure for clustering. The algorithm can obtain more accurate and stable clustering result by deleting the outliers before electing the initial centroids.

VI.CONCLUSION

Thus, the project is used to handle the large amount of data using map reduce techniques.Clustering algorithms are used to find the patterns in data and these clustering algorithms must also scale well with the increasing amount of data. In this project using k-means clustering algorithm can use to cluster the massive dataset calculate the SSE and find the accuracy, making use different type of dataset in the Hadoop framework cluster performance. MapReduce manages the data partitions and carries on parallel processing on the portioned data. Cluster analysis based upon distance metric selected and the criterion for determining the order of clustering. Depending upon the centroids may yield different results. The number of iterations required for cluster data can be reduced by choosing the right set of initial set of centroids.

REFERENCES

[1] DweepnaGarg, KhushbooTrivedi, B. B. Panchal , "A Comparative study of Clustering Algorithms using MapReduce in Hadoop", *International Journal of Engineering Research & Technology*, Vol. 2.

[2] Prajesh P Anchalia, Anjan K Koundinya, Srinath N K, "MapReduce Design of K-Means Clustering Algorithm*", IEEE International Conference*, 2013.

[3] Fahim, A. M., A. M. Salem, F. A. Torkey, and M. A. Ramadan , "An efficient enhanced k-means clustering algorithm", Journal of Zhejiang University Science, Vol. 7 No. 10 pp.1626-1633, 2006.

[4] Dean, Jeffrey, and Sanjay Ghemawat, "MapReduce: simplified data processing on large clusters", Communications of the ACM, Vol. 51, No. 1 pp. 107-113, 2008.

[5] Slagter, Kenn, Ching-Hsien Hsu, Yeh-Ching Chung, and Daqiang Zhang, "An improved partitioning mechanism for optimizing massive data analysis using MapReduce", The Journal of Supercomputing, Vol.66, No.1, pp. 539-555, 2013.

[6] Wang, Juntao, and Xiaolong Su, "An improved K-Means clustering algorithm", In procIEEE 3rd International Conf on Communication Software and Networks (ICCSN), Vol. 30, No.7, pp. 44-46, 2011.

[7] Chen, Min, Shiwen Mao, and Yunhao Liu (2014), "Big Data: A Survey", Mobile Networks and Applications, Vol.19 No. 2 pp.171-209, 2014.

[8] N. KarthikeyaniVisalakshi and J. Suguna, "K-Means Clustering using Max-min Distance Measure", *IEEE International conference*, 2009.

[9] Raed T. Aldahdooh, WesamAshour, "Distance-based Initialization Method for K-means Clustering Algorithm", *IEEE conference*, vol 02, pp 41-51, 2013.

[10] M. Halkidi, Y. Batistakis, M. Vazirgiannis, "Clustering algorithms and validity measures", *International conference on information and computer Networks*, pp 3-22, 2001.