Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 6.017

IJCSMC, Vol. 6, Issue. 8, August 2017, pg.84 – 88

Map Reduce Design for Data Clustering

Priyanka A
Computer Science & VTU, India
priya.bhat.a@gmail.com

Abstract— The data is of no use unless or until the useful information is retrieved from it. It is impractical and inefficient to use traditional database techniques since the data is generated in practical scenario is huge in volume and it is exchanged in faster way than ever before. That is why a new open source framework called Hadoop comes in to existence. Hadoop is used to process huge amount of data called big data. To extract useful information and to derive some meaningful conclusions from massive amount of data, datamining techniques can be used. Among many datamining techniques, clustering is most popular one. Clustering is a process of binding data members with similar characteristics in to one group whereas dissimilar data members are distributed into different groups. In this paper K-Means clustering algorithm is implemented using map reduce programing model. The key idea behind the KMeans implementation is design of mapper and reducer routines. The technique presented here asks to enter the number of clusters that is K and large data file. It finds centroid for each cluster and distance from each item to centroid of the cluster is computed. The data is assigned to the group with which the distance of the data is least.

Keywords— Hadoop, Data Mining, Big Data, K-Means, Clustering

I. INTRODUCTION

In today's world, the production of data which is coming quickly from various sources creates difficulty in their collection, storage, analysis, management and processing. The generated information becomes useless unless or otherwise they are analysed correctly. Big data provides a set of techniques to handle these challenges in order to grab the opportunities that they offer.

Data mining is one such technique used to discover useful information from huge data set. But existing data mining techniques cannot be applied directly on huge volume of data because it increases the complexity of processing. Clustering is one such approach in data mining used to analyse large volume of data generated by real world applications. Clustering is a process of partitioning data in such a way that data with similar characteristics grouped into single cluster and data with dissimilar characteristics belong to different clusters. In traditional method this process becomes tedious as the data increases in quantity. There are different types of clustering algorithm used to partition data in to clusters.

A. Clustering Algorithms

1) Hierarchical Algorithms: In this algorithm data is ordered depending on the method of nearness (proximity).The proximity is obtained by the middle nodes. A dataset is represented by dendrogram, where every information represented by leaf nodes. As the hierarchy continues the initial group (cluster) is progressively divided in to several groups.

2) Partitioning Algorithms: These algorithms divide datasets into a different number of partitions, in which every partition is considered as a single cluster. The following requirements must be satisfied by these clusters. Each cluster should contain minimum of one data object. Each data object must belong to only one cluster.

3) Density-Based Algorithms: In this algorithm, based on the region of density and connectivity the data items are grouped into clusters. They are closely linked with point –nearest neighbours.

4) Grid-Based Algorithms: In this data object space is divided in to grids. This algorithm scans the input dataset only once to find statistical value for grids. Based on accumulated grid data grid based clustering is applied.

5) Model-Based Algorithms: In this method clustering is done based on some already defined mathematical model. This approach is depending on an assumption that data is generated by a combination of probability distribution. This method automatically determines the number of clusters based on statistical data.

In this paper, KMeans clustering algorithm is implemented for large data set. The KMeans clustering algorithm comes under the category of partition based algorithm. The K-Means algorithm is an unsupervised algorithm used to group dataset with similar characteristic into one cluster and data with dissimilar properties belong to other cluster. This algorithm is used to extract some meaningful information from the available dataset. It allocates n objects to k partitions in such a way that every data item belong to nearby partition.

The main requirement for efficient clustering method is

- Intra-clustering similarity should be maximum
- Inter clustering similarity should be minimum

B. About Hadoop

To implement KMeans clustering on big data an open source java framework called Hadoop is used. Hadoop framework works based on Map Reduce programming model, which consists of mapper and reducer task.

•The Mapper Task: This task is the one which takes input data and changes it into a set of data, where every single component is broken down into tuples of the form key-value pairs.

•The Reducer Task: The output of the mapper task is given as an input to this task. It combines the data tuples into a tuples of smaller set. This task is executed after the execution of mapper task.

II. RELATED WORK

Purnawansyah and Haviluddin [3] in this paper they proposes KMeans clustering implementation in network traffic activities. Here, k-means was used to identify bandwidth used by the users per day. The algorithm was applied on 456 data set and grouping is done based on low, medium and high bandwidth usage. This algorithm was implemented using MATLAB2013 tool. The information analysed in this paper was information exchange over a period of 30 minutes per day for three units (Rectorat, science, forestry) in five months. The data was normalized before sending through network between 0.05 and 1.The normalized value was calculated for the data using original, minimum, maximum of the raw data.

QingHe, XinJinChangyingDu, FuzhenZhuang and Zhongzhi Shi [4] tells about Clustering in extreme learning machine feature space in which ELM K-Means and ELM NMF uses ELM feature on Kmeans to solve the clustering problem. It was a partition based algorithm which takes dataset from document corpus as well as from UCI machine learning repository. If the number of nodes was more than 300 than this algorithm gives very good accuracy else optimal performance cannot be obtained.

R Madhuri, M Ramakrishna Murty, JVR Murthy, PVGD Prasad Reddy and Suresh C Satapathy [5] says Cluster analysis on different datasets using K-modes and K-prototype algorithms. This was a kind of partitioning based algorithm and takes data set from Mixed Numeric and Categorical data .The quality of cluster generated by applying this method is not very good.

III. EXISING SYSTEM

Clustering is an important technique in data mining to retrieve a meaningful information. There are different types of clustering algorithms existing which are applied on different types of datasets. These techniques uses traditional database management tools to store the data. The traditional storage systems can store only some limited amount of data. As the data size increases data access and management becomes a tedious task. When the huge amount of data is given to these algorithm, the time it takes to process these data sets also increases and cluster quality reduces. So we need to have a tool to process massive volume of data.

IV. PROPOSED SYSTEM

The graphical representation of the proposed system is shown below.

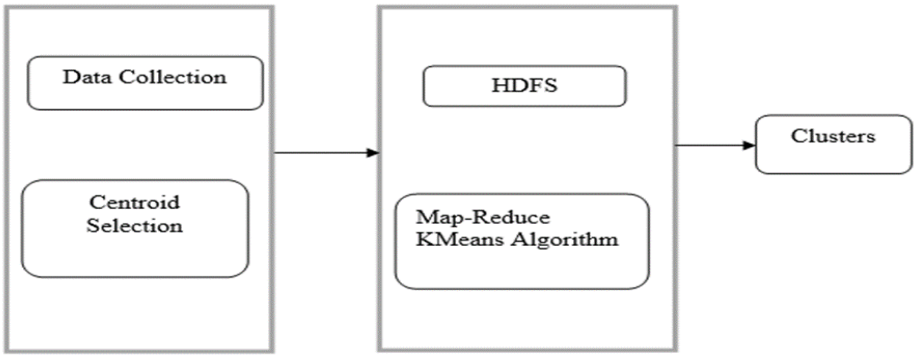

Fig. 1 Model of Proposed System

Proposed system uses an open source framework called Hadoop. It consists of HDFS and map reduce programming model. Since HDFS uses data replication concept, the loss of data is impossible. This file system is capable of managing large amount of data, data access becomes easier. In this project KMeans clustering algorithm is implemented using map reduce construct. The mapper function computes Euclidean distance between cluster centroid and data point and the reducer function recalculates the cluster centroid. So computation time it takes to group the data with similar characteristic in to a same cluster is reduced and cluster quality is improved.

V. IMPLEMENTATION

In this section map reduce methods are implemented using Hadoop framework to cluster huge amount of data set using KMeans clustering algorithm.

A. Algorithm for map reduce KMeans

- Read the data set
- Compute the vector value for each data point and store it in a vector file
- Input number of clusters
- Select the cluster centroids randomly and store it on a cluster file
- Run the Hadoop components
- Copy the vector and cluster file onto HDFS file system
- Run MapReduce K-Means algorithm
- Copy the cluster centroids obtained in the last iteration into local file system
- Based on cluster centroids assign the data into different cluster

B. Mapper Function

This method takes input as a list of <key, value >pair where key is the location of the data object and value is its content.

This method produces output in the form of <key, value>pair in which key is index of the cluster and value is the data object belonging to that cluster. This function performs following task.

- Calculate the distance between data object and each cluster centroids
- Assign the data object to its nearest cluster centroid
- Repeat above two steps until all data objects are processed

C. Reducer Function

It takes input as a list of < key, value > pair. The output produced by this function is of the form <key, value > pair where key represents the index of the cluster and value specifies the centroid for the new cluster. This function performs following tasks.

- Computes sum of data objects in each cluster
- Computes number of data objects in each cluster
- Computes the mean of data objects belonging to cluster. This is called as new cluster centroid
- Repeat the above steps until the termination condition reaches.

The dataset is collected from UCI Machine Learning Dataset Repository. This dataset consists 20000 text documents from 20 different newsgroups.

VI. RESULTS

Before executing Kmeans algorithm, it is essential to check Hadoop components are running successfully. This is done by executing jps command. In case if any component is not started successfully, then necessary steps has to be taken for successful running of individual Hadoop component.

Fig. 2 Check all Components of Hadoop are running

The output of map reduce program is stored in a default file called part-r-00000.This file is present in HDFS. Here, part-r-00000 contains centroids of each cluster after 3 iterations.

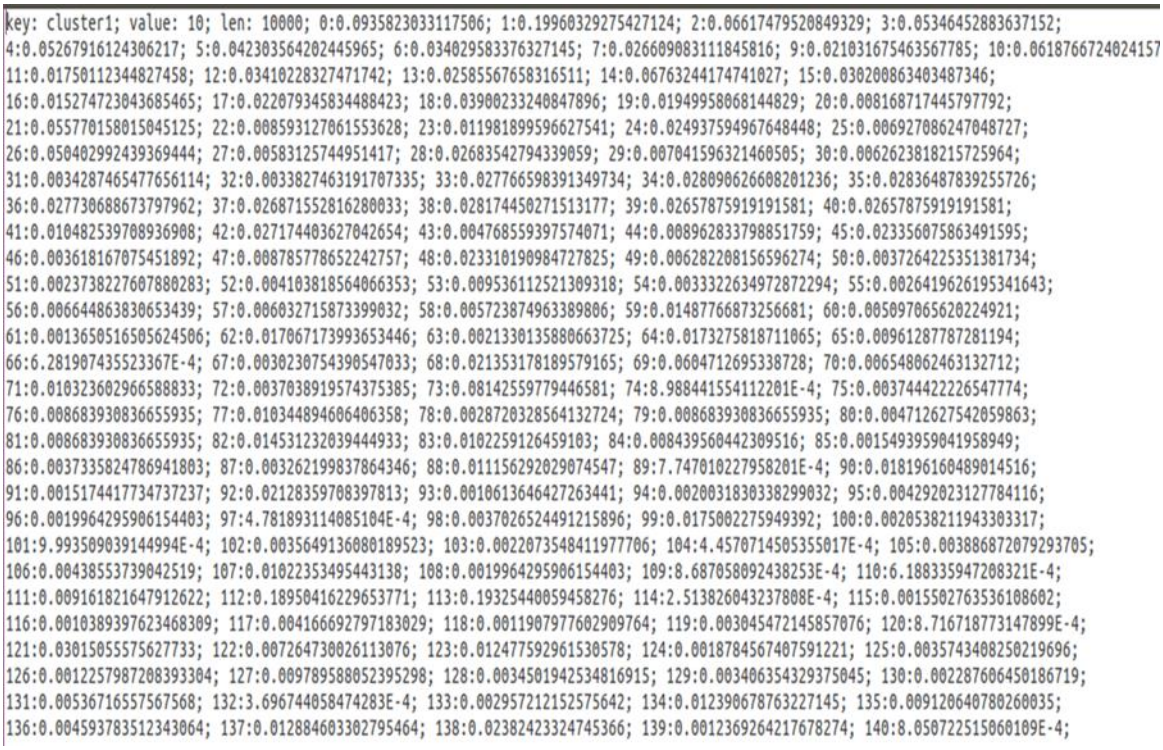

Fig. 3 Part-r-00000 file

Here text files are grouped based on final centroids obtained as a result of the execution of map reduce programs. The output is 20 clusters with number of text documents from each subdirectory assigned to each cluster.

VII. CONCLUSIONS

The amount of data exchange in these days requires high quantity of data processing. This project implements K-Means Algorithm in a single node cluster using Hadoop framework to group given text documents based on certain patterns. This application provides an efficient method and robust system to group data with similar characteristics, decreases the problems in handling huge volume of data and also reduces the execution time.

REFERENCES

[1] P. P. Anchalia, A. K. Koundinya and S. N. K., "*MapReduce Design of K-Means clustering Algorithm*", International Conference on Information Science and Applications (ICISA), Suwon, pp. 1-5,2013.
[2] Adil Fahad,Najlaa Alshatri,Zahir Tari, Abdullah Alamri, "*A Survey of clustering algorithms for big data*", IEEE, Vol 2, pp 267-279,September 2014.
[3] Purnawansyah and Haviluddin,"*K-Means clustering implementation in network traffic activities*", IEEE, International Conference on Computational Intelligence and Cybernetics, Makassar, pp. 51-54, 2016.
[4] Qing He, Xin Jin, Changying Du, Fuzhen Zhuang, and Zhongzhi Shi.,"*Clustering in extreme learning machine feature space.Neurocomputing*",pp 88-95, 2014.
[5] R Madhuri, M Ramakrishna Murty, JVR Murthy, PVGD Prasad Reddy, and Suresh C Satapathy, "*Cluster analysis on different data sets using k-modes and k-prototype algorithms*", In ICT and Critical Infrastructure: Proceedings of the 48th Annual Convention of Computer Society of India-Vol II, pp 137-144, Springer, 2014.
[6] Apache Hadoop: http://hadoop.apache.org/
[7] www.bogotobogo.com/Hadoop/BigData_hadoop
[8] www.michael-noll.com/tutorials

