Density Based Clustering and Fuzzy Clustering for Efficient Clustering of Big Data in Hadoop Ecosystem

Lingaraj K[1*], Priyanka A[1], Jerabandi preethi[1] and Syeda shameemunnisa.[1]

Abstract Map Reduce is a parallel technology which can process bigdata by creating multiple instances of thread. Map will take input data and split it into multiple chunks or parts and distributes all parts to different reducers. All reducer will process the data and send result back to Mapper. Mapper gathers output from all mappers and then generates a single output. Due to multiple parallel processing of Map Reduce technology allow us to process any amount of data. we discuss clustering algorithms such Density Based and Fuzzy Clustering. Both algorithms are not efficient to group all similar data to single cluster and make some data to compromise by putting little un- similar to different clusters. To implement this project author is using National Climatic Data Center (NCDC) dataset which contains climate information. To find out similar temperature on different dates author is applying Hybrid clustering algorithm. From this dataset author is using date and temperature value and then passing date as key to Mapper and temperature as value to Mapper.

1 Introduction

Big data as of now is exploring in the market and information is quickly moving from gigabytes to zeta bytes [1]. Huge information has such huge information pre requisites those applications that were recently used to store and procedure information-Database Management System (DBMS), Relational Database Management System (RDBMS) and so forth. There are presently bombing the information demand [2]. Huge information incorporates incredibly enormous data sets implying that it isn't workable for usually utilized programming apparatuses to oversee and process that information inside the necessary time frame [3]. Along these lines, greatly equal programming stumbling into numerous servers is presently required to deal with this work load [4]. Large information expects strategies to uncover bits of knowledge from datasets that are various, complex, and of an enormous scope. A portion of the difficulties of large information handling remember challenges for information catch, addressing the requirement for speed, tending to information quality, managing exceptions, sharing of enormous information, and huge information analysis[5]. Various strategies have been proposed to information so as to deal with enormous information datasets, e.g., AI, affiliation systems, bolster vector machines, and clustering [6]. Right now, propose cross breed grouping strategy to deal with large information.

Lingaraj K
e-mail: lingaraj.k10@gmail.com

Priyanka A
e-mail: ankireddypriyankacse@gmail.com

Jerabandi preethi
e-mail: jerabandipreethi@gmail.com

Syeda Shameemunnisa
e-mail: syedashameems996@gmail.com

[1] Dept. of Computer Science and Engineering, Rao Bahadur Y Mahabaleshwarappa Engineering College, Bellary, Karnataka – 583104, India

2 Related Work

2.1 K-Means Clustering Algorithm

The diverse research papers using the K-implies grouping for the enormous information are extravagantly talked about beneath, Sreedhar C etal. proposed a K-Means Hadoop Map Reduce (KM- HMR) for the successful large information bunching. Right now were introduced for Map Reduce (MR) system based bunching [4]. The KMHMR was the primary strategy, which focused on the use of MR on standard K-implies. The subsequent strategy was to improve the groups quality by limiting the between bunch removes and expanding intra-bunch separations. The proposed KM-HMR approaches results have out flanked the productivity of other grouping techniques regarding execution time Nadeem Akthar et al. prescribed a changed K-means grouping calculation, which chooses the K-ideal information focuses in the dataset. The principle bit of leeway of choosing the information focuses from the enormous datasets is to forestall exception focuses from including in the last assessment of the group [5]. Increasingly steady outcomes were achieved when the underlying focuses were arranged for proper data sets. Prasanta K. Jana proposed a K- implies grouping calculation executed in Spark. The proposed K-Means calculation tackled the goals issues which is available in the normal K-Means bunching calculation by earlier re-optimization of the in for groups [6]. It brought about better execution of the Spark structure based K-implies bunching calculation even with the expanded size of the information and even machine check.

2.2 Varient of K-Means Clustering Algorithm

The distinctive research papers using the Variant of K-Means bunching for the huge information are intricately discussed about beneath, Mohamed Aymen Ben HajKacem et al. proposed the Accelerated Map-Reduce-based K-Prototypes (AMRKP) grouping procedure for taking care of large information. Right now perusing and composing of information were done just once on account of this the quantity of Input and Output (I/O) tasks were decreased radically [7]. Further most, the proposed plot is reliant on pruning system to quicken the way toward bunching by minimization the excess separation between the inside and information purposes of the group. The created AMRKP outperform the other grouping plans regarding productivity and adaptability. Omair Shafiq proposed an equal K-Medoids grouping calculation dependent on MR structure for doing the compelling bunching of enormous database. The contrived bunching technique was proficient, straightforward and equipped for dealing with the datasets with fluctuating differing attributes, similar to volume, speed and assortment. The reenactment result showed the capacity and attainability of proposed bunching technique in taking care of enormous scope datasets [8]. Mohamed Aymen Ben Haj Kacem et al. proposed a MR system utilizing K-Prototypes (MR-KP) bunching plan for the powerful information grouping utilizing parallelization. It brought about being a popular and viable bunching technique for blended gigantic datasets. There enactment was performed on a huge number of tests and the result was exact and versatile in any event, when the size of information is expanded [9].

2.3 Fuzzy C-Means Clustering Algorithm

The diverse research papers using the FCM bunching for the huge information are intricately examined beneath; Simone A Ludwig explored the adaptability and parallelization of FCM grouping calculation. The FCM bunching calculation was parallelized utilizing MR structure by illustrating the methodology of guide and lessens work. The approval investigation of the MR-FCM bunching calculation was made to show the adequacy of the proposed calculation as further

part of virtue [10]. Minyar Sassi Hidri etal. proposed an upgraded FCM grouping calculation utilizing inspecting blended in with part and consolidation procedure for bunching large information . Starting advance is to part information into unmistakable subsets and work the individual hubs in parallel. At that point, the subsets were inspected, which were again part haphazardly into particular sub-samples. This calculation performed adequately with the furnished assets with upgraded existence complexities [11].

2.4 Possiblistics C-Implies (PCM) grouping Algorithm

The diverse research papers using the PCM bunching for the large information are intricately talked about underneath, Qingchen Zhang and Zhikui Chen recommended a weighted piece PCM calculation (wKPCM) for grouping the information objects in to the reasonable gatherings. The part loads were coordinated for characterizing the item's significance in the bit grouping for limiting the defilement created by the uproarious information. The proposed dispersed wKPCM grouping calculation depended on the MR system which can prepare convincing computational speed for continuous informational collections [12]. Qingchen Zhang etal. proposed a Privacy Preserving High-Order PCM (PPHOPCM) grouping calculation for playing out the bunching of huge information through the upgrading the goal work. The disseminated HOPCM technique depends on the MR structure with the point of managing large information [13]. In the end, the PPHOPCM was utilized for ensuring the information on cloud by applying the Brakerski-Gentry-Vaikuntanathan (BGV) encryption plot on HOPCM .In the PPHOPCM, participation framework and group focuses were refreshed utilizing polynomial capacities. The formulated PPHOPCM calculation adequately grouped the gigantic dataset and secures the private cloud information.

2.5 Collabrative Filtering (CF) based bunching algorithm

The diverse research papers using CF bunching for the huge information are extravagantly examined underneath, Rong Huet al. planned a Clustering-based Collaborative Filtering (CF) approach whose intension is to offer comparative types of assistance enlistment in similar groups for the suggestion of administrations synergistic. This methodology of grouping is included two stages. In the main stage the informational indexes are disintegrated into little pieces of bunches to make them reasonable text preparing [14]. In the following stage CF is applied to the decided groups. Since the number administrations include in the bunch was not exactly the accessible web benefits the time intricacy of CF was lesser nearly. Subramaniya swamy V. et al. proposed the prescient component based CF technique for the successful handling of enormous scope information parallel. The MR system is utilized for doing the accumulation, sifting and upkeep of the proficient stockpiling. The CF was accustomed to refining of information [15]. The created grouping plan was improved by preparing the information into emoji and tokens through the utilization of notion investigation. The reenactment brought about critical improvement in multifaceted nature investigation execution.

3 Density Based Clustering Algorithm

Cluster analysis is a data mining technique that searches, recommends and organizes the data. In the technique of clustering, datasets are assembled into various numbers of clusters with different attributes [7, 8]. Clustering belongs to unsupervised learning techniques, unlike classification, in which alike objects of the dataset are aggregated into clusters[9], and thus form unique clusters in which objects that belong to same cluster are very different from each other and objects in the

same group or cluster are very similar to each other [10, 11]. We get to know the clusters only after the complete execution of clustering algorithm [12]. To manage large datasets we use Density Based and Fuzzy clustering algorithm which are explained below. We are finding clusters using Density Based algorithm and then applying Fuzzy clustering with Map-Reduce on density based cluster to find compromise points and put them in similar cluster.

Density based clustering giving 30% precision and after applying hybrid algorithm with density and fuzzy giving 100% precision result. Lower the precision value lower the quality of cluster and higher the precision higher the quality of cluster.

Precision = Total_clusters_find_out /expected clusters (3.1)

Density Based Clustering: The DBSCAN algorithm is based on this intuitive notion of "clusters" and "noise". The key idea is that for each point of a cluster, the neighborhood of a given radius has to contain at least a minimum number of points. Fuzzy Clustering: Fuzzy clustering (also referred to as soft clustering or soft k-means) is a form of clustering in which each data point can belong to more than one cluster.

4 Experimental Results

In this paper, simulation is carried out for NCDC data Set. The simulation was performed using Cloudra software were 2020 year dataset of different cities contains temperature and date values are uploaded.

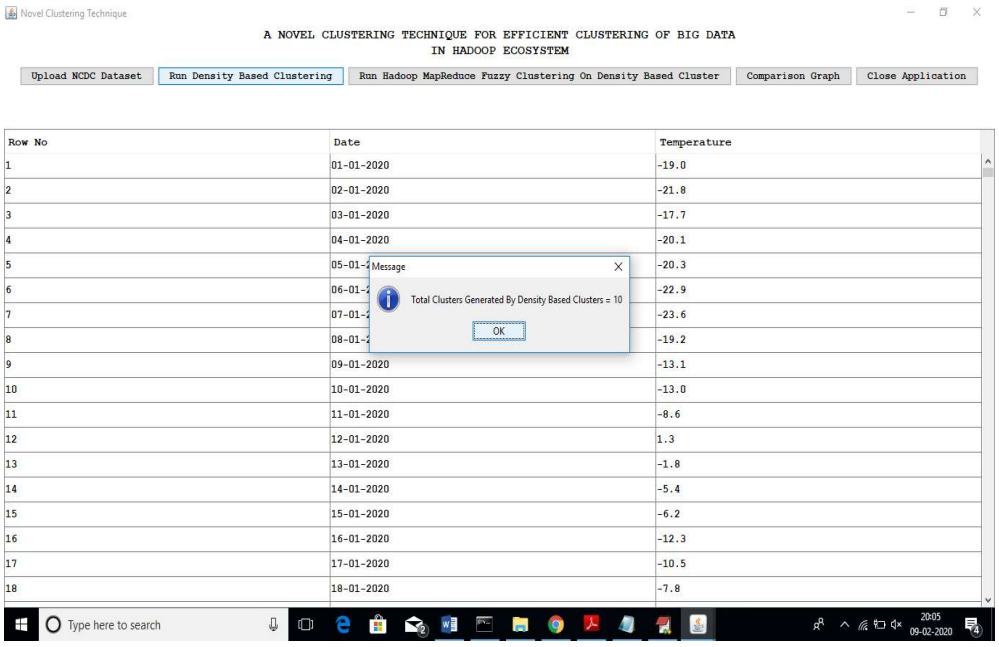

Fig. 1. In above screen message we can see density based created to total 10

In fig 2 we can see all 10 clusters and now we know density based put some related closed points in different clusters due to that reason we got 10 clusters. Now to reduce clusters size and to put related data in similar cluster we apply Fuzzy algorithm with Map-Reduce (also called as Hybrid Algorithm)

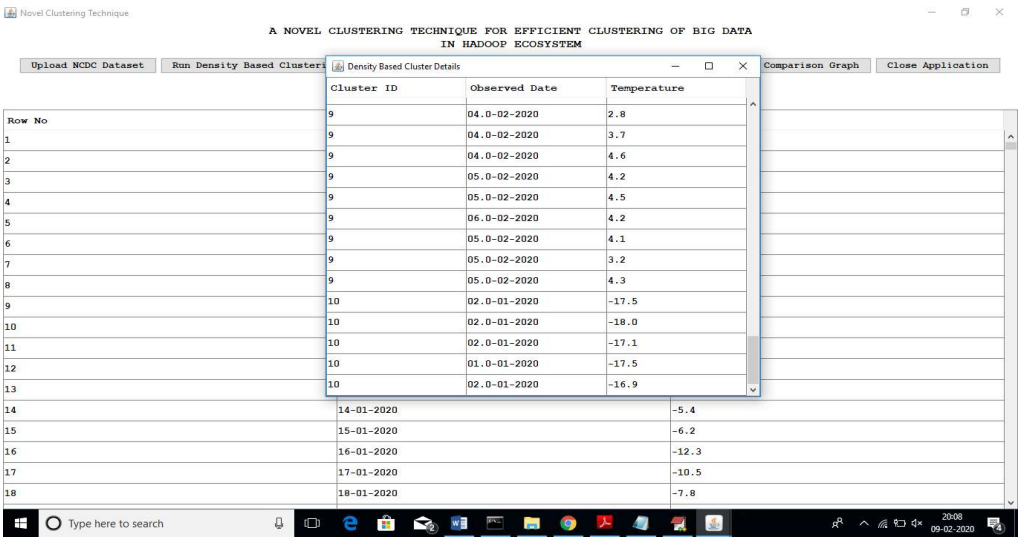

Fig. 2. In above screen message hybrided algorithm of density based clustering

In above graph x-axis represents algorithm name and y-axis represents precision value. From above graph we can see density based got 30% precision value and Hybrid Map-Reduce based with fuzzy and density based combination got 100% precision. In below screen we can see Map-Reduce processing details of data splits.

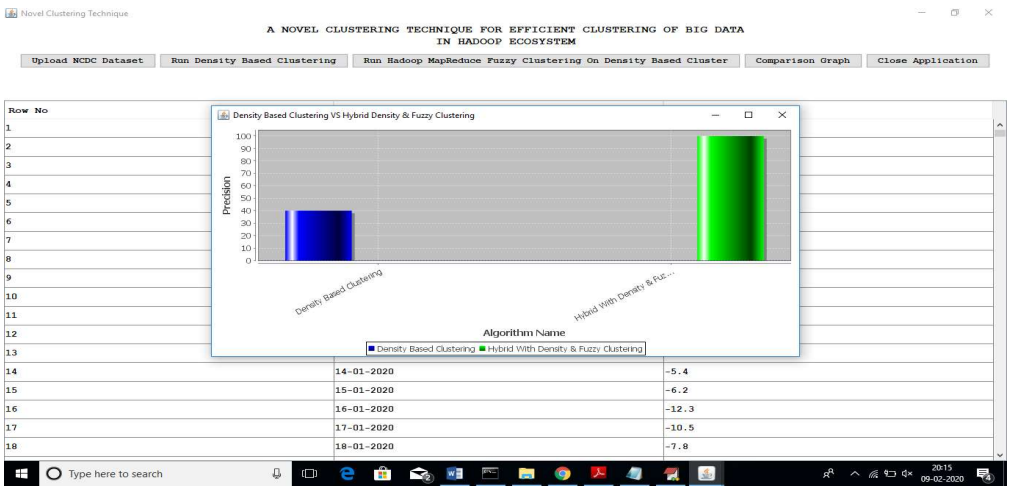

Fig. 3. In above screen message shows precision

5 Conclusion

Map Reduce is a parallel technology which can process bigdata by creating multiple instances of thread. Map will take input data and splitit into multiple chunks or parts and distribute all parts to different reducers. All reducers will process the data and send result back to mapper. Mapper gather output from all mappers and then generate a single output. Due to multiple parallel processing of MapReduce technology allow usto process any amount of data. Wehave introduced a Density Based and Fuzzy clustering algorithm using execution times than those of the Density Based Clustering: The DBSCAN algorithm is based on this intuitive notion of “clusters” and “noise”. The key idea is that for each point of a cluster, the neighborhood of a given radius has to contain at least a minimum number of points.

References

[1] Wullianallur Raghupathi, and Viju Raghupathi, "Big data analytics in healthcare:promiseandpotential", Health information science and systems,vol.2,no.1,pp.3,2014.

[2] Bhagyashri S. Gandhi, and Leena A. Deshpande, "The survey on approaches to efficient clustering and classification analysis of big data", International Journal of Engineering Trends and Technology (IJETT),vol.36,no.1,pp. 33-39, 2016.

[3] Ali Seyed Shirkhorshidi, Saeed Aghabozorgi, Teh Ying Wah and Tutut Herawan, "Big Data Clustering: A Review", Computational science and its applications - ICCSA 2014: 14th international conference Guimarães, Portugal, june30-july3, 2014 proceedings.

[4] Chowdam Sreedhar, Nagulapally Kasiviswanath,and Pakanti Chenna Reddy, "Clustering large datasets using K-means modified inter and intra clustering (KM-I2C) in Hadoop", Journal of Big Data ,vol.4, no.1,pp.27,2017.

[5] Nadeem Akthar, Mohd Vasim Ahamad and Shahbaz Khan, "Clustering on Big Data Using Hadoop MapReduce", in proceedings of 2015 IEEE International Conference on Computational Intelligence and Communication Networks (CICN), pp. 789-795,2015.

[6] Ankita Sinha, and Prasanta K. Jana, "A novel K-means based clustering algorithm for bigdata", In proceedings of 2016 IEEE International Conference on Advances in Computing, Communications and Informatics (ICACCI),pp.1875-1879,2016.

[7] Mohamed Aymen Ben Haj Kacem, Chiheb-Eddine BenN’cir, and Nadia Essoussi, "One-pass Map Reduce- based clustering method for mixed large scale data", Journal of Intelligent Information Systems, pp.1-18,2017.

[8] M. Omair Shafiq, and Eric Torunski, "A Parallel K-Medoids Algorithm for Clustering based on Map Reduce", in proceedings of 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA),pp.502- 507,2016.

[9] Mohamed Aymen Ben Haj Kacem, Chiheb-Eddine BenN'cir, and Nadia Essoussi, "Map Reduce-based k- prototypes clustering method for big data", in proceedings of 2015 IEEE International Conference on Data Science and Advanced Analytics(DSAA),pp.1-7,2015.

[10] Simone ALudwig, "Map Reduce- based fuzzy c-means clustering algorithm: implementation and scalability" ,International Journal of Machine Learning and Cybernetics, vol.6,no.6,pp.923-934,2015.

[11] Minyar Sassi Hidri, Mohamed Ali Zoghlami, and Rahma Ben Ayed, "Speeding up the large-scale consensus fuzzy clusteringfor handling Big Data", Fuzzy Sets and Systems, 2017.

[12] Qingchen Zhang, and Zhikui Chen, "A weighted kernel possibilistic c-means algorithm based on cloud computing for clustering big data", International Journal of Communication Systems, vol. 27, no. 9, pp. 1378- 1391,2014.