Challenges in Big Data Clustering for Data Analytics

Cowsalya T, Karthikeyan N,
Asst. Professor,
Dept of Computer science and Engineering,
SVS College of Engineering,
Coimbatore.
narenkarthikeyan.mecse@gmail.com

Gayathri S, Priya A
UG Scholar,
Dept of Computer science and Engineering,
SVS College of Engineering,
Coimbatore,

***Abstract*-- Big Data is regularly defined by its three characteristics such as Volume, Variety and Velocity (3V's). It refers to the data that are too big, vibrant and composite. In this perspective, data are complex to capture, manage, analyze, and store using traditional database management tools. So, the new conditions forced by Big Data present serious challenges at different level, including data clustering. This paper aims to give a preface to cluster analysis, and then focus on the challenge of clustering big data.**

***Index Terms*-- Cluster, Map Reduce, Data mining and K means.**

I. *INTRODUCTION*

Cluster analysis divides data into clusters (groups) for the purposes of summarization or better understanding. For instance, cluster analysis has been used to collect related information for browsing or as a means of data compression. Although clustering has a extended history and a huge number of clustering techniques have been introduced in statistics, data mining, pattern recognition, and other fields, major challenges still remain. The term "clustering" is used in several research communities to describe methods for grouping of unlabeled data. Clustering [1-3] is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data. The essential requirements of clustering algorithms are scalability, ability to deal with noisy data, insensitive to the order of input records, etc. Data mining is a multi-step process. It requires collecting and converging data for a data mining algorithm, prospecting the data, evaluating the results and taking relevant action. The collected data can be stored in one or more operational databases, a data warehouse or a flat file. The data is mined using two learning approaches. i.e. supervised and unsupervised clustering. Data Mining is a four step process: Assemble data, Apply data mining tools on datasets, Interpretation and evaluation of result, Result application.

To be sure, the tremendous challenge of Big Data is to make diverse data (weather forecasting, logistics, geographically located data, and traffic data) and to associate them to take out useful information and thus improve the various sectors exploit this enormous amount of data very wide and discrete. According to Heterogeneous, Autonomous, Complexity, Evolving (HACE) theorem [5] the most important characteristics of Big Data are Heterogeneous data, Autonomous Complexity and Evolution.

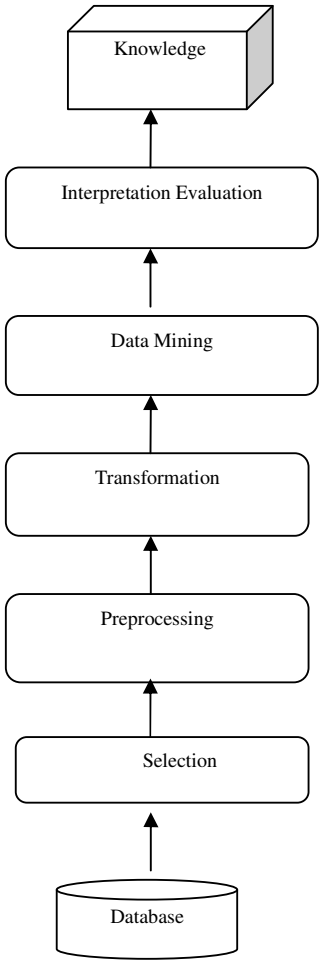

Figure 1: Steps of Data Mining Process

In Heterogeneous data, data comes from several different sources like Twitter, Facebook, LinkedIn and instant messaging in complex and heterogeneous format which requires a set of techniques and the implementation of various solutions. In autonomous, depending on autonomous sources gives Big Data one of its main characteristics. In this sense, this source consists on distributed and decentralized controls. so each data sources can work independently without being based on any centralized control. The same principle is found in World Wide Web (WWW) setting in which each web server is capable to generate the information and to function correctly

without involving other servers. On the other hand, the complexity of Big Data makes her very vulnerable so it will easily malfunction if it were relaying on any centralized control unit. Another point is that having autonomous servers helps some Big Data applications like Google or some social networks (facebook) to provide quick responses and nonstop services for clients.

Complexity of Big Data is linked to multiple data; the data is collected in very different contexts (multi-source, multi-view, multi-tables, sequential, etc.) as well as decentralized treatment data or massively parallel processing (MapReduce). Data complexity increases with the increase in volume and the usual treatment methods, with management of relational database tools are no longer sufficient to meet the requirements capture, storage and further analysis. The evolution of complex data also represents an essential feature. Big data is changing very quickly. The typical example is when a customer commented on a page of social networking, these comments must be extracted over periods of a specific time so that the algorithm can operate and have relevant information

II. BIG DATA CLUSTERING TECHNIQUES

Generally, Big Data clustering techniques can be classified into two categories [6]: single machine clustering techniques and multiple machine clustering techniques, recently the latter draws more attention because they are faster and more adapt to the new challenges of Big Data, single-machine techniques and clustering multiple machines include different techniques as is illustrated in Figure 2.

A. Single-machine clustering

Datamining clustering algorithms: The unsupervised classification (clustering) is an essential datamining tool for the analysis of Big Data, which aims to consolidate the significant class data objects (clusters) so that objects grouped in the same cluster are similar and consistent according to specific parameters. It is difficult to apply data mining clustering techniques in Big Data because of the new challenges. So with the great mass of data provided by the Big Data and the complexity of clustering algorithms which have very high treatment costs, the question that arises is how to deal with this problem and how to deploy clustering techniques Big Data to obtain results in a reasonable time. The complexity of Big Data makes her very vulnerable so it will easily malfunction if it were relaying on any centralized control unit.

B. Partitioning based clustering algorithms:

This method divides a data set in a single partition using a distance to classify points based on their similarities; the drawback of the partitioning methods is that those methods generally require the user to a predefined K parameter for a clustering solution which is often non-deterministic. There are many partitioning algorithms such as K-means [7], k-medoids K-modes, PAM, CLARA, CLARANS and FCM.

C. Hierarchical based clustering algorithms:

This method partitions data into different levels that resemble a hierarchy. This classification provides a clear data visualization. The aim of this method is to collect objects into classes increasingly wide, using some measures of similarity or distance. The results of this type of classification are usually represented as a tree of hierarchical classification. The hierarchical method has a major drawback, which is related to the fact that once a stage is completed, it cannot undone. BIRCH, CURE, ROCK and Chameleon are some algorithms well known in this category.

D. Density based clustering algorithms:

The clustering approach based on density [12, 13] is able to find clusters in an arbitrary manner, where the clusters are defined as dense regions separated by low density areas, generally, clustering algorithms based on density are not suitable for large data sets, DBSCAN, OPTICAL DBCLASD and DENCLUE are the algorithms using this method to filter noise (outliers).. *E. Model based clustering algorithms:*

Clustering algorithms is worked based on a mixture model we can measure the uncertainty of the classification by a law of multivariate probability distributions where each mixture represents a different cluster, the classification problem based on a model is that the processing time is very slow in case of large data sets.

F. Grid based clustering algorithms:

It refers three stages: firstly is to divide the space into rectangular cells to obtain a grid of cells of equal size, and then delete the low density of cells, and finally combine adjacent cells having a high density to form clusters. The great advantage of grid-based classification is the significant reduction in complexity. Some examples are: GRIDCLUS, STING, CLICK and WaveCluster. The major limitation of clustering algorithms is their instability, which means that implementing the same algorithm can provide different results from one moment to another.

G. Dimension Reduction:

The data size can be measured in two dimensions, the number of variables and the number of examples. These two dimensions can take very high values, which could cause a problem during the exploration and analysis of these data. For this, it is essential to implement data processing tools and make a pretreatment to the dataset before applying clustering algorithms for a better understanding of the value of knowledge available in this data. The Dimension reduction technique is one of the oldest approaches to provide answers to this problem. Its purpose is to select or extract optimal subset of relevant features for a criteria already fixed. The selection of this subset of features can eliminate irrelevant and redundant information according to the criterion used. This selection or extraction makes it possible to reduce the size of the sample space and makes it all more representative of the problem. For large sets of data, dimension reduction is usually performed before applying the classification algorithm to avoid the disadvantages of high dimensionality. It refers feature selection and feature extraction. In feature selection, it aims to select an

optimal subset of variables from a set of original variables, according to a certain performance criteria. The main objective of this selection is to reduce the number of required actions. Feature extraction aims to select features in a transformed space - in a projection space, the extraction methods use all the information to compress and produce a vector of smaller dimension.

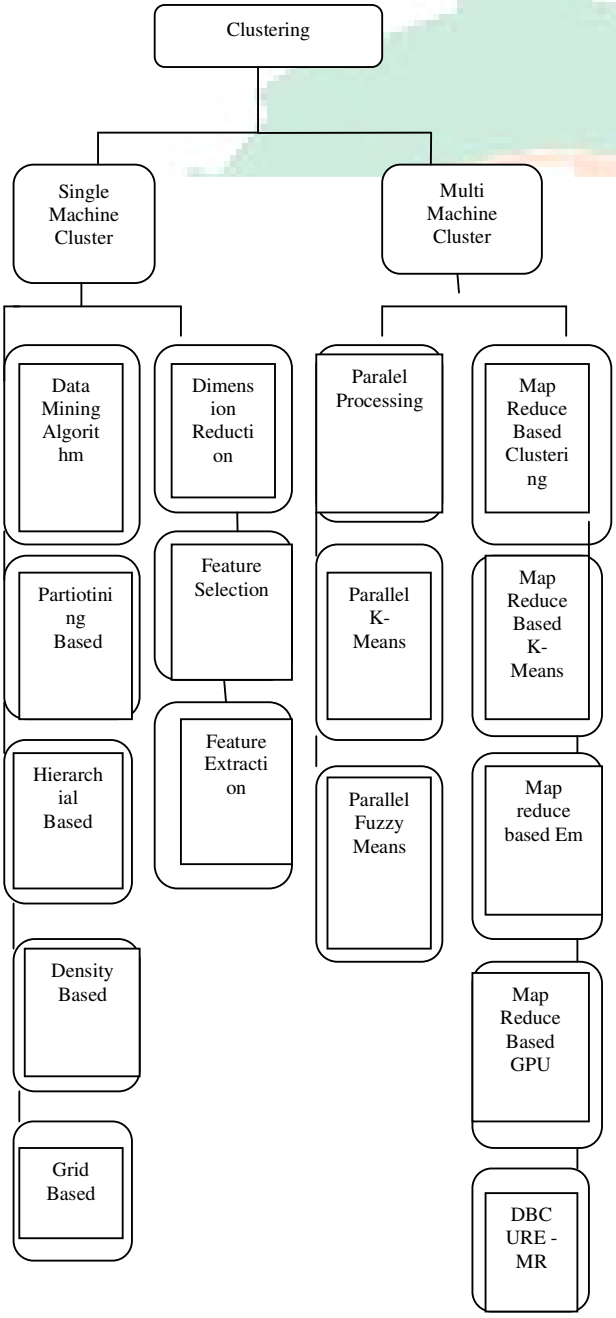

Fig. 3. Clustering architecture and it types

III. MULTI MACHINE CLUSTERING

A. Parallel clustering:

The processing of large amounts of data imposes a parallel computing to achieve results in reasonable time. In this section, we examine some parallel algorithms and distributed clustering used to treat Big Data, the parallel classification divides the data partitions that will be distributed on different machines. This makes an individual classification to speed up the calculation and increases scalability.

B. MapReduce based clustering:

MapReduce is a task partitioning mechanism (with large volumes of data) for a distributed execution on a large number of servers. Principle is to decompose a task (the map part) into smaller tasks [8-10]. The tasks are then dispatched to different servers, and the results are collected and consolidated (the reduce part). There are several approximate methods that have used this framework to improve existing clustering algorithms. An approach to accelerate the K-means clustering method is used.

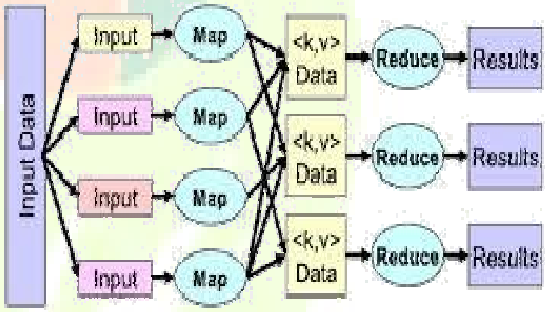

Fig. 4. MapReduce Function

On the other hand, another study addresses the Big Data processing problem using the K-means algorithm that proposes a new model of treatment with MapReduce to eliminate iteration dependency and achieve high performance. Another parallel method is proposed by adapting the EM algorithm in MapReduce, so that the main memory in each computer just needs to load a set of data. This method can reduce the time and memory

IV. CONCLUSION

Clustering lies at the heart of data analysis and data mining applications. The ability to discover highly correlated regions of objects when their number becomes very large is highly desirable, as data sets grow and their properties and data interrelationships change. At the same time, it is notable that any clustering "is a division of the objects into groups based on a set of rules. It is neither true nor false. Clustering can be done by the different number of algorithms such as hierarchical, partitioning, grid and density based algorithms. Hierarchical clustering is the connectivity based clustering. Partitioning is the centroid based clustering [11]. However, the

MapReduce framework can provides a very good basis for the implementation of such parallel algorithms. Generally, in order to manage large volume of data while keeping an acceptable resource needs, we have to improve clustering algorithms by reducing, their complexity in terms of time and memory.

REFERENCES

[1] K.Kameshwaran and K.Malarvizhi, " Survey on Clustering Techniques in Data Mining," International Journal of Computer Science and Information Technologies, Vol. 5 (2) , pp.2272-2276, 2014,

[2] A. Sherin, S. Uma, K.Saranya and M. Saranya Vani, "Survey On Big Data Mining Platforms, Algorithms And Challenges". International Journal of Computer Science & Engineering Technology,Vol. 5 No, 2014.

[3] S. Arora and I. Chana, ""A survey of clustering techniques for Big Data analysis,"" in Confluence The Next Generation Information Technology Summit (Confluence), 5th International Conference-. IEEE, p. 59-65, 2014.

[4] M. Tyagi, F. Bovolo, A. K. Mehra, S. Chaudhuri and L. Bruzzone, "A Context-Sensitive Clustering Technique Based on Graph-Cut Initialization and Expectation-Maximization Algorithm", IEEE Geoscience and Remote Sensing Letters, Vol. 5(1), pp. 21 - 25, 2008.

[5] F. Bu, Z. Chen, Q. Zhang, and X. Wang, "Incomplete Big Data Clustering Algorithm Using Feature Selection and Partial Distance," In Digital Home (ICDH), 5th International Conference on. IEEE, p. 263-266, 2014.

[6] M. G. Vadgasiya and J. M. Jagani, " An enhanced algorithm for improved cluster generation to remove outliers ratio for large datasets in data mining,"" Inter. J. Advance Eng. and Research, vol-1, pp-203-208, 2014.

[7] Shuo Chen and Chengjun Liu, "Clustering-Based Discriminant Analysis for Eye Detection", IEEE Transactions on Image Processing, vol. 23(4), pp. 1629-1638, 2014.

[8] Jong Soo Park; Ming-Syan Chen; P. S. Yu, "Using a hash-based method with transaction trimming for mining association rules," IEEE Transactions on Knowledge and Data Engineering, vol. 9(5), pp. 813-825, 1997.

[9] M. J. Mendenhall; E. Merenyi, "Relevance-Based Feature Extraction for Hyper spectral Images," IEEE Transactions on Neural Networks, vol. 19(4),pp. 658-672, 2008.

[10] R. D. Dony and S. Haykin, "Neural network approaches to image compression" IEEE Proceedings, vol. 83(2), pp. 288-303, 1995.

[11] C. H. Wu, C. S. Ouyang, L. W. Chen and L. W. Lu, "A New Fuzzy Clustering Validity Index With a Median Factor for Centroid-Based Clustering", IEEE Transactions on Fuzzy Systems, vol. 23(3), pp. 701-718, 2015.

[12] P. Batra nagbal, and P. Ahlawat mann, "Survey of Density Based Clustering Algorithms," International journal of Computer Science and its Applications, vol. 1, no 1, pp. 317,2011

[13] R. Xu and D. Wunsch, ""Survey of clustering algorithms," Neural Networks, IEEE Transactions, vol. 16, no 3, pp. 645-678 2005.