International Journal of Engineering and Techniques - Volume 4 Issue 4, July - Aug 2018

RESEARCH ARTICLE OPEN ACCESS

Algorithms and Challenges in Big Data Clustering

R.Suganya[1], M. Pavithra[2], P.Nandhini[3].

[1](Assistant Professor, Department of C.S.E, Jansons Institute of Technology, Coimbatore, India)

[2](Assistant Professor, Department of C.S.E, Jansons Institute of Technology, Coimbatore, India)

[3](PG Scholar, Department of C.S.E, Jansons Institute of Technology, Coimbatore, India)

Abstract:

Big data is data sets that are so big and complex that traditional data-processing application software are inadequate to deal with them.Big Data is usually defined by three characteristics called 3Vs (Volume, Velocity and Variety). It refers to data that are too large, dynamic and complex. In this context, data are difficult to capture, store, manage, and analyze using traditional data management tools. Thus, the new conditions imposed by Big Data present serious challenges at different level, including data clustering. This paper aims to review the state of art and make a concise synthesis related to clustering techniques in Big Data context.

Keywords — Big Data, Clustering algorithms,MapReduce, Data Mining, Dimension reduction, Parallel classification.

I. INTRODUCTION

BIG Data refers to a very strong growth of heterogeneous data flows due to the increased use of new technologies. In fact, with the growth of the web, the use of social networks, mobile, connected and communicating objects, information is now more abundant than ever and it is growing faster every day [6]. Some studies argue that handling and using intelligently this huge data could become a new pillar of economics as well as scientific research, experimentation and simulation [2]. Indeed, many opportunities of Big Data appear in different areas such as health (enhancing the efficiency of some treatments), biomedical, marketing (increasing sales), transportation (reducing costs), business, finance (minimizing risks), management (decision making with high efficiency and speed), social, media, and government services [3]. In this paper, we introduce the most popular Big Data's clustering techniques: single machine clustering techniques and multiple machine clustering techniques, including Data mining clustering algorithms, dimension reduction techniques, parallel classification and the MapReduce framework. Most state of the art papers found in the literature focus on a single category of clustering techniques whereas our goal here is to make a broad and general synthesis concerning the Big Data clustering issues and pinpoint the advantages of the important techniques [4]. The paper is organized as follows: the second section represents the related work which shows different state of the art papers. The third section describes the challenges of Big Data. The fourth section provides a global view of the various clustering techniques dealing with Big Data's

challenges and showing how to exploit a large amount of data [5].

II. RELATED WORK

Zomaya et al. [1] present a survey of existing clustering algorithms of different categories (Partitioning-based, Hierarchical-based, Density-based, grid-based and modelbased). In their work they established a comparison between five categories with their most representative algorithm; their goal was to find the best performing for Big Data. In [2] the authors focus on the most popular and most used algorithms in the literature like k-means, they presents some comparative work of these algorithms. Another recent research [3] presents a general view of data mining algorithms and platforms that can be used in the field of Big Data by discussing different challenges and characteristics. Paper [4] discusses some of Big Data mining algorithms to find the most appropriate among them using a comprehensive comparison. Nagpal and Mann's paper [5] does not address all the clustering technique it is interested only to study density based clustering algorithms such as DBSCAN DENCLUE and to discuss their advantages and disadvantages. Others in [6] are interested in studying classification algorithms that can be used in statistics and apply them to specific databases. Researchers in [7] present a review of some old algorithms that can handle large data set as Nearest Neighbor Search, Decision Tree and Neural Network. In [8], Herawan et al. discuss different clustering techniques including MapReduce, parallel classification using MapReduce. They present an overview of different categories of data mining clustering algorithms. Our study covers all the above techniques. It deals with different categories of data mining clustering algorithms and discusses their advantages and disadvantages. We also present techniques that can be used to deal with new requirements imposed in Big Data context such as dimension reduction, parallel classification, MapReduce framework. We will conclude with a comparison between all these techniques.

III. BIG DATA CHALLENGES

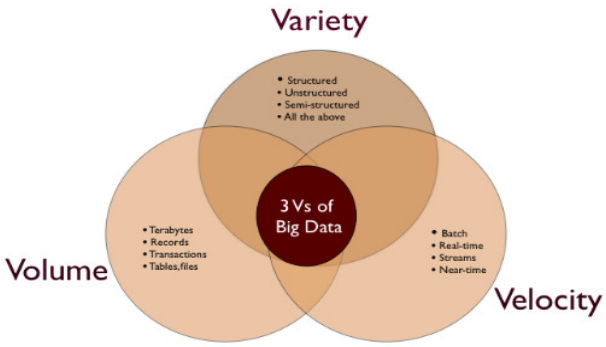
The mass of data available on the web grows exponentially. In general, we find that the data can be classified into three basic types: structured (are the basic data types such as integers, characters, and arrays of integers or characters. They are used in relational databases), unstructured (have no predefined format: email, books, journals, documents, videos, photos) and semi-structured (are a combination of two previous types of data, they are generally represented using XML) [6]. Most of produced data are unstructured and traditional database management tools are unable to handle this type of information [3].

The 3Vs defining Big Data are:

•**Volume:** today, masses of data to be processed are ever increasing. This describes the fact that our increased use of new technologies (smartphones, social networks, connected machines ...) encourages to produce more and more data in our daily activities both personal and professional; the companies are facing an explosion of stored data [7]. Indeed, this volume continues to grow at high speed. It is estimated that the volume of data stored in the world doubles every four years. It has stored more data since 2010 than it had been done since the beginning of humanity [9].

•**Velocity:** The notion of velocity of Big Data refers to the speed at which data is generated, captured and exchanged. Indeed, these data are generated and evolve very rapidly [2]. So the collection, analysis and use of data should more often be done in real time, it is even possible to

stop storing information and analyzing flow (streaming), to draw the right conclusions [4].



•**Variety:** The last "V" is made that the data is very varied and has not always structured forms. Indeed, It can use the data contained in websites, blogs, emails, exchanges on social networks (Facebook, Twitter, LinkedIn ...), images, video, audio, logs, data spatial (geolocation), the biometrics, etc. Their origins are diverse: web, text mining, mining picture, etc [1]. We need to combine several sources to draw actionable conclusions. The variety of Big Data explains the difficulty of using the information from traditional data warehousing infrastructure. Indeed, the extreme challenge of Big Data is to make heterogeneous data (weather, logistics, geolocation, car traffic) and associate them to extract useful information and thus improve the various sectors exploiting this huge amount of data very wide and dispersed [3]. According to HACE (Heterogeneous, Autonomous, Complexity, and Evolving) theorem [9] the most important characteristics of Big Data.

•**Heterogeneous data**, that's means that data comes from several different sources like Twitter, Facebook, LinkedIn and instant messaging in complex andheterogeneous format which requires a set of techniques and the implementation of various solutions [6].

•**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 [7]. 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 [9].

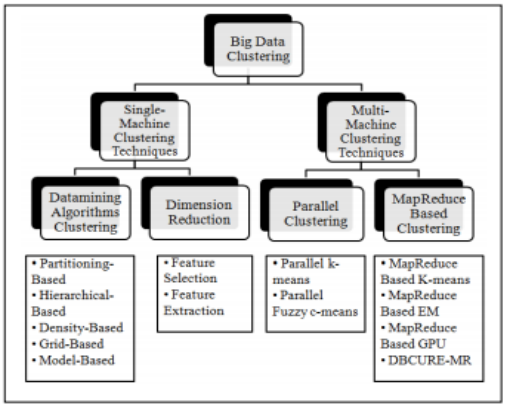
• **Complexity**, the complexity of Big Data is linked to multiple data; the data is collected in very different contexts (multi-source, multi-view, multi-tables, sequential, etc.) as well as decentralized treatment data or massively parallel processing (MapReduce) [8]. Data complexity increases with 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.

•**Evolving**, 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 [2]. To manage the growing demands of data, we should increase the capacity and performance of tools and methods. Big Data requires new solutions to improve the capacity and effective treatment to exploit functionally of the data without necessarily recruit new resources [3]. Indeed, with the exponential growth of data, traditional data

mining algorithms have been unable to meet important needs in terms of data processing. So in order to exploit this huge amount of data, efficient processing model with a reasonable computational cost of this huge, complex, dynamic and heterogeneous data is needed [4].

IV. BIG DATA CLUSTERING

Generally, Big Data clustering techniques can be classified into two categories [8]: 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.



A. Single-machine clustering

1) Data mining clustering algorithms:

The unsupervised classification (clustering) is an essential data mining 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 [12]. 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 [13]. There are many different classification methods in the literature. These methods can be classified into: partitioning methods, hierarchical methods, methods based on a grid, density-based methods and methods based on a model, this taxonomy is inspired from articles of state of art in the field [11].

a) 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 [14]. There are many partitioning algorithms such as K-means, k-medoids K-modes, PAM, CLARA, CLARANS and FCM.

b) 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 [10]. The results of this type of classification areusually 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 undo. BIRCH, CURE, ROCK and Chameleon are some algorithms well known in this category [3].

c) Density based clustering algorithms:

The clustering approach based on density 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) [4].

d) Model based clustering algorithms:

with the clustering algorithms 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, examples of this type of classification algorithms are EM, COBWEB, CLASSIT and SOM [5].

e) Grid based clustering algorithms:

Complys with the 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 gridbased classification is the significant reduction in complexity. Some examples are: GRIDCLUS, STING, CLICK and WaveCluster [6].

Cls. Algorithms	*Size of Data*	*Cls. Quality*	*Scalability*	*Stability*
EM	Large	High	Low	Suffer from
FCM	Large	High	Low	Suffer from
DENCLUE	Huge	Partially	High	Suffer from
OptiGrid	Huge	Partially	High	Suffer from
BIRCH	Huge	Partially	High	Suffer from

The algorithm BIRCH, DENCLUE and OptiGrid are more adapted to large amounts of data but they suffer from low classification quality.

2) 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 [7]. 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 [3]. 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 [8].

a) 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. A work made in 2014 [12] proposes a classification algorithm for Big Data based on feature selection. Firstly, the feature selection algorithm is designed to reduce the size of the dataset. Then, a parallel k-means algorithm is applied to the data subsets selected in the first step. Experimental results show that the proposed algorithm provides better classification accuracy than existing algorithms and takes much less time than other classification algorithms for Big Data [11].

b) Feature extraction:

It 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. A work performed in 2012 [13] proposes a classification algorithm and feature extraction for Big Data that is based on PCA and LS-SVM, the experimental results on the Big Data shows that the proposed classification algorithm based on feature extraction allows solving large classification problems [12].

B. Multiple-machine clustering:

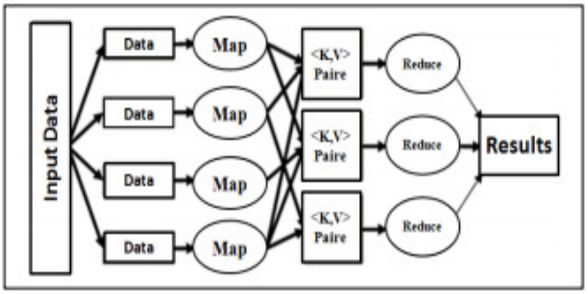
1) Parallel clustering:

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 [5]. This makes an individual classification to speed up the calculation and increases scalability.

A parallel k-means algorithm was proposed [4], which was then implemented on an IBM SP2 POWER parallel with 16 nodes. In the other hand, [5] have implemented a further parallel version of the k-means algorithm using 32 machines on an Ethernet network and which showed an almost linear acceleration for large data sets. The scalability of the parallel k-means algorithm has also been demonstrated by others[7].

2) 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 [6]. The tasks are then dispatched to different servers, and the results are collected and consolidated (the reduce part).The function of this framework is shown in Figure.



In step Map the input data is analyzed, cut into subproblems and delegated to other nodes (which can do the same recursively). This will be processed later using the Map function which has a pair (key, value) that associates a set of new pairs (key, value) [3]. Then comes the stage Reduce, where the lowest nodes reach their results back to the parent node that had asked them. It calculates a partial result using the Reduce function (reduction) involving all the corresponding values for the same key to a unique pair (key, value). Then he goes back information in turn. There are several approximate methods that have used this framework to improve existing clustering algorithms [5].

An approach to accelerate the K-means clustering method is proposed [8]. Later, [9] improves the idea of accelerating the K-means algorithm. It proposed a fast K-means algorithm that gives a better approximate solution. 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 [1]. Another parallel method [2] 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. Younghoon and Kyuseok develop the DBCURE-MR [3]. It is the DBCURE algorithm parallelized using MapReduce.

While traditional algorithms based on a density found each cluster one by one, the DBCURE-MR found several clusters in parallel. Experimental results confirm that DBCURE-MR found the clusters effectively. Other studies [4] show how the GPU and the Dynamic Parallelism feature of CUDA platform can bring significant benefits to BIRCH, the GPU can accelerate BIRCH and make up to 154 times faster than the CPU version with good scalability and high precision [8]. Here below is a recapitulative table including different techniques of clustering in terms of Big Data, their advantages and strength as well as their limitations [9].

Clustering Techniques	Advantages	Limitations
Datamining clustering algorithms	-Simple implementation	-Don't have the capacity to deal with huge amount of data
Dimension reduction	-Reduce the dataset -Optimize treatment cost -Very fast and scales algorithm	-Don't offer an efficient solution for high dimensional datasets -Should be performed before applying the classification algorithm
Parallel classification	-Minimize the time of execution -More scalable	-Implementing the algorithms can't easily be done
MapReduce framework	-Offer impressive scalability - Generate instance responses -Inherently Parallel	-Need more resources -Implementing each query as a MR program is difficult -No primitives for common operations(selection/extraction)

CONCLUSION

This paper describes different methodologies and different algorithms used to manage large sets of data. It shows that these algorithms are insufficient to face all the challengesraised by the Big Data. Indeed there is no clustering algorithm that can be used to solve all the Big Data issues [10]. Although the parallel classification is potentially very useful for Big Data clustering, but the complexity of the implementation of these algorithms remains a great challenge [8]. 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 [5].

REFERENCES

[1] A. Fahad, N. Alshatri, Z. Tari, A. ALAmri, A. Y. Zomaya, I. Khalil, F. Sebti, and A. Bouras, "A Survey of Clustering Algorithms for Big Data: Taxonomy & Empirical Analysis," IEEE transactions on emerging topics in computing, 2014.

[2] A.benayed, M.benhalimaand M. alimi, "Survey on clustering methods: Towards fuzzy clustering for Big Data," In Soft Computing and Pattern Recognition (SoCPaR), 6th International Conference of. IEEE, p. 331-336, 2014.

[3] A. Sherin, S. Uma, K.Saranya and M. SaranyaVani"Survey On Big Data Mining Platforms, Algorithms And Challenges". International Journal of Computer Science & Engineering Technology,Vol. 5 No, 2014. [4] S.ARORA, 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.

[5] P. BatraNagpal, and P. Ahlawat Mann, "Survey of Density Based Clustering Algorithms," International journal of Computer Science and its Applications, vol. 1, no 1, p. 313-317,2011

[6] R. Xu and D. Wunsch, "Survey of clustering algorithms," Neural Networks, IEEE Transactions, vol. 16, no 3, p. 645-678, 2015.

[7] C. Yadav, S. Wang, et M. Kumar, "Algorithm and approaches to handle large Data-A Survey," International Journal of computer science and network, vol 2, issue 3, 2013.

[8] A. S. Shirkhorshidi, S. Aghabozorgi, T. Y. Wah, and T. Herawan, "Big Data Clustering: A Review," In Computational Science and Its Applications–ICCSA 2014. Springer International Publishing, p. 707- 720. 2014.

[9] X. Wu, X. Zhu, G. Q. Wu, and W. Ding, "Data mining with Big Data," Knowledge and Data Engineering, IEEE Transactions on, vol. 26, no 1, p. 97-107, 2014.

[10] C.C. Aggarwal, C.K. Reddy, Data Classification: Algorithms and Applications. CRC Press, 2014.

[11] M. G. Vadgasiya and J. M. Jagani, "An enhanced algorithm for improved cluster generation to remove outlier's ratio for large datasets in data mining," Development, vol. 1, no 11, 2014.

[12] 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.

[13] B. J. Kim, "A Classifier for Big Data," In Convergence and Hybrid Information Technology. Springer Berlin Heidelberg, p. 505-512, 2012.

[14] I. S. Dhillon, and D.S. Modha, "A data-clustering algorithm on distributed memory multiprocessors," In Large-Scale Parallel Data Mining. Springer Berlin Heidelberg, p. 245-260, 2010.

[15] K. Stoffel and A. Belkoniene, "Parallel k/h-means clustering for large data sets," In Euro-Par'2009 Parallel Processing. Springer Berlin Heidelberg, p. 1451-1454, 2009.