ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

A Survey on Analysis of Big Data Clustering Techniques and Challenges

Ch. Radhika

Asst. Professor, Dept of CSE, G. Narayanamma Institute of Technology and Science, Hyderabad, India

ABSTRACT: 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 database management tools. There are various techniques to analyse Big data. As Big data varies from terabytes to petabytes of data which leads to high computational costs and complexity. The question is how to cope with this problem and how to deploy a suitable technique to assess Big data and to get the results in a reasonable time. This paper aims to review and make a concise survey related to clustering techniques in Big data context.

KEYWORDS: Big data; clustering; parallel; MapReduce

I. Introduction

Big data comprises massive sensor data, raw and semi-structured log data of IT industries and the exploded quantity of data from social media. Examples of this data include high-volume sensor data and social networking information from web sites such as Google, Face Book, LinkedIn, Yahoo, Amazon and Twitter. 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. The exponential growth of data in all fields demands the revolutionary measures required for managing and accessing such data. Big data need big storage and this volume makes operations such as analytical operations, process operations, retrieval operations very difficult and time consuming. One way to overcome these difficult problems is to have big data clustered in a compact format. Clustering is the task of grouping input data into subsets called clusters. Data clustering is a well-known technique in various areas of computer science and related domains. Although data mining can be considered as the main origin of clustering, but it is vastly used in other fields of study such as bio informatics, energy studies, machine learning, networking, pattern recognition.

Traditional clustering techniques cannot cope with this huge amount of data because of their high complexity and computational cost. The main target is to scale up and speed up clustering algorithms with minimum sacrifice to the clustering quality. Big data clustering techniques can be classified into two categories single machine clustering techniques and multiple-machine clustering techniques.

In this paper, the introduction of 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 clustering and the MapReduce based clustering are covered. The goal here is to make a broad and general synthesis concerning the Big data clustering issues and pinpoint the advantages of the important techniques. The paper is organized as follows: The second 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. The last section presents the conclusion and possible improvements of clustering techniques.

II. CLUSTERING TECHNIQUES WITH BIG DATA

Generally, Big data clustering techniques can be classified into two categories [1]: 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 [8] of Big data. As it is demonstrated in Fig. 1.

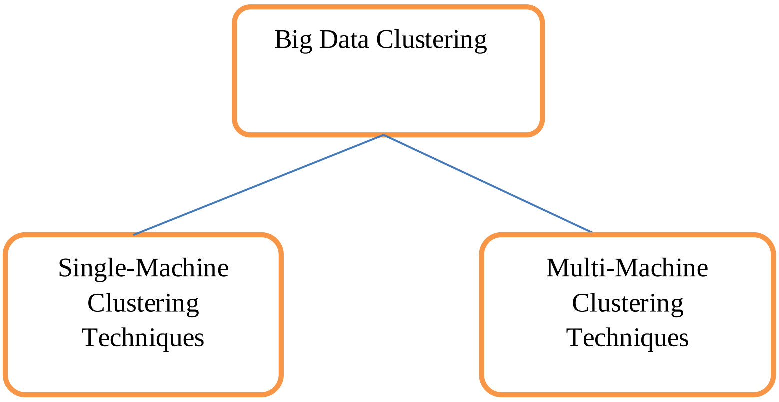
Big Data Clustering
Single-Machine Clustering Techniques
Multi-Machine Clustering Techniques

Fig.1. Big data clustering techniques

The Clustering Techniques are compared using the following factors [9]-

1. **Dataset size-** It means the volume or the size of the dataset; it can be small, medium or large.
2. **Dataset type-** It means the type of the attributes in the dataset. It can be numeric, categorical, mixed etc.
3. **Cluster shape-** A good clustering algorithm should be able to produce clusters of **arbitrary** shapes.
4. **Time complexity-** Time for obtaining the final clusters. A successful algorithm should take lesser time to cluster a large volume of data.
5. **Handle outlier-** Outliers are the points containing false information that make it difficult for an algorithm to cluster the data into the suitable or true cluster. Hence they should be properly handled or removed.

A. Single machine clustering techniques

Single-machine clustering techniques [4] run in one machine and can use resources of just one single machine.

As it is illustrated in Fig.2.

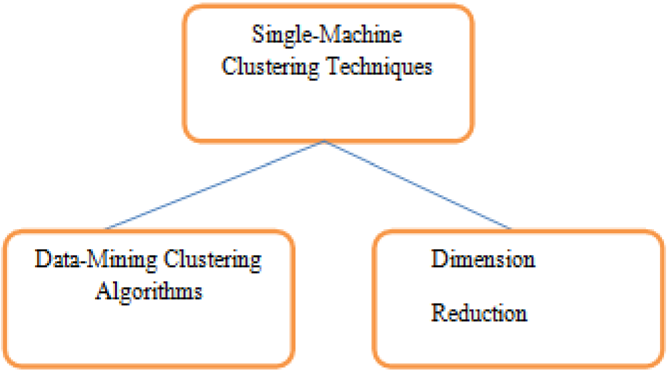
Single-Machine Clustering Techniques
Data-Mining Clustering Algorithms
Dimension Reduction

Fig.2. Single machine clustering techniques

A. I. Data-Mining Clustering Algorithms

Partitioning-based: The partitioning algorithms divide data objects into a number of partitions, where each partition represents a cluster. There are many partitioning algorithms [13] such as K-means, k-medoids, K-modes, PAM, CLARA, CLARANS and FCM.

Challenges:

-Poor at handling noisy data and outliers.
-Works only on numeric data.
-Empty cluster generation problem.
-Random initial cluster center problem.
-User has to provide the value of k

Hierarchical-based: This method partitions data into different levels that resemble a hierarchy. Hierarchical clustering [13] methods can be agglomerative (bottom-up) or divisive (top-down). BIRCH, CURE, ROCK and Chameleon are some of the well-known algorithms of this category.

Challenges:

-If an operation (merge or split) is performed, it cannot be undone i.e. no backtracking is possible.
-Inability to scale well.
-It is order-sensitive and may generate different clusters for different orders of the same input data.
- May not work well when clusters are not spherical.

Density-based: Here, data objects are separated based on their regions of density, connectivity and boundary. They are closely related to point-nearest neighbors. A cluster, defined as a connected dense component, grows in any direction that density leads to. DBSCAN, OPTICS, DBCLASD and DENCLUE [11] are algorithms that use such a method to filter out noise (outliers) and discover clusters of arbitrary shape.

Challenges:

- Unsuitable for high-dimensional datasets due to the curse of dimensionality phenomenon.

-Its quality depends upon the threshold set.

Grid-based: The space of the data objects is divided into grids [11]. The main advantage of this approach is its fast processing time, because it goes through the dataset once to compute the statistical values for the grids.

Some examples are: GRIDCLUS, STING, CLICK and Wave-Cluster.

Challenge:

-Depends only on the number of cells in each dimension in the quantized space.

Model-based: Model based [4] clustering method optimizes the fit between the given data and some (predefined) mathematical model.

Examples of this type of classification algorithms are EM, COBWEB, and CLASSIT.

Challenges:

- The processing time is very slow in case of large data sets.
- Complex in nature.

III. DIMENSION REDUCTION TECHNIQUES

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 [11] is usually performed before applying the classification algorithm to avoid the disadvantages of high dimensionality.

Challenges:

-Don’t offer an efficient solution for high dimensional datasets.

-Should be performed before applying the classification algorithm.

B. Multi-Machine clustering techniques

Nowadays the growth of data size is very much faster than memory and processor advancements, consequently one machine with a single processor and a memory cannot handle terabytes and petabytes of data and it underlines the need algorithms that can be run on multiple machines. The multiple-machine clustering techniques [3] can run in several machines and has access to more resources. The division is shown in Fig.3.

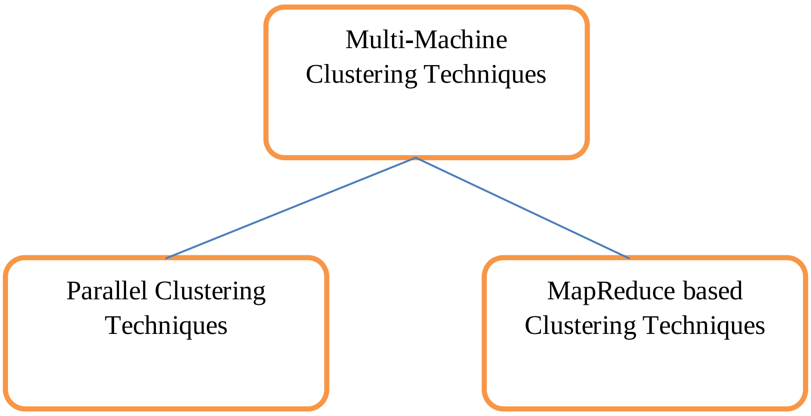
Multi-Machine Clustering Techniques
Parallel Clustering Techniques
MapReduce based Clustering Techniques

Fig.3. Multi machine clustering techniques

Multi machine clustering algorithms are divided into two main categories:

- Un-automated distributing– parallel
- Automated distributing– MapReduce

I. Parallel clustering

In parallel clustering [6], developers are involved with not just parallel clustering challenges, but also with details in data distribution process between different machines available in the network as well, which makes it very complicated and time consuming. The processing of large amounts of data imposes a parallel computing to achieve results in reasonable time. 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. A parallel and distributed clustering algorithm follows a general cycle.

In the first stage, data is going to be divided into partitions and they distribute over machines. Afterward, each machine performs clustering individually on the assigned partition of data. Two main challenges for parallel and distributed clustering [12] are minimizing data traffic and its lower accuracy in comparison with its serial equivalent. Lower accuracy in distributed algorithms could be caused by two main reasons, first, it is possible that different clustering algorithms deploy in different machines and secondly even if the same clustering algorithm is used in all machines, in some cases the divided data might change the final result of clustering.

Although parallel algorithms [7] add difficulty of distribution for programmers, but it is worth full because of the major improvements in scaling and speed of clustering algorithms.

Challenges:

-Complexity of implementing the algorithms can’t be done easily.

IV. MAPREDUCE BASED CLUSTERING

Although parallel clustering algorithms improved the scalability and speed of clustering algorithms still the complexity of dealing with memory and processor distribution was a quiet important challenge. Difference between parallel algorithms and the MapReduce framework is in the comfortless that MapReduce[10] provides for programmers and reveals them form unnecessary networking problems and concepts such as load balancing, data distribution, fault tolerance and etc. by handling them automatically. This feature allows huge parallelism and easier and faster scalability of the parallel system. MapReduce is a task partitioning mechanism for a distributed execution [12] on a large number of servers. Principle is to decompose a task (the map part) into smaller tasks. The tasks are then dispatched to different servers, and the results are collected and consolidated (the reduce part).

Challenges:

-Need more resources.

-Implementing each query as a MR program is difficult.

-No primitives for common operations (selection/extraction).

V. CONCLUSION

This paper describes the different clustering techniques and the algorithms with the challenges they pose with Big data. Parallel clustering is very useful for big data clustering, but the complexity of implementation is a great challenge. However, the MapReduce framework can provides a very good basis for the implementation of such parallel clustering.

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. 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. Btissam Zerhari, Ayoub Ait Lahcen and Salma Mouline, "Big Data Clustering: Algorithms and Challenges", International Conference on Big Data, Cloud and Applications BDCA'15 , At Tetuan, Morocco , conference paper may 2015.

3. 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.

4. Yasodha P, Ananathanarayanan NR. "Analyzing Big Data to build knowledge based system for early detection of ovarian cancer." Indian Journal of Science and Technology. 2015 Jul; 8(14):1–7.

5. Pandove D, Goel S. "A comprehensive study on clustering approaches for Big Data mining". IEEE Transactions on Electronics and Communication System; Coimbatore. 2015 Feb 26-27. p. 1333–8.

6. Btissam Zerhari, Ayoub Ait Lahcen and Salma Mouline, "Big Data Clustering: Algorithms And Challenges", International Conference on Big Data, Cloud and Applications BDCA'15, At Tetuan, Morocco, conference paper may 2015.

7. Apurva Juyal Dr. O. P. Gupta,"A Review on Clustering Techniques in Data Mining",International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 7, July 2014.

8. T. Sajana, C. M. Sheela Rani and K. V. Narayana," A Survey on Clustering Techniques for Big Data Mining" Indian Journal of Science and Technology, Vol 9(3), DOI:10.17485/ijst/2016/v9i3/75971, January 2016.

9. Ali Seyed Shirkhorshidi , Saeed Aghabozorgi , Teh Ying Wah and Tutut Herawan, "Big Data Clustering:A Review",2015.

10. S.M. Junaid, K.V. Bhosle," Overview of Clustering Techniques", International Journal of Advanced Research in Computer Science and Software Engineering,Volume 4, Issue 11, November 2014.

11. 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.

12. Justin Samuel, Koundinya RVP, KothaSashidhar and C.R. Bharathi, A Survey on Big Data and its Research Challenges, ARPN Journal of Engineering and Applied Sciences, Vol. 10, No. 8, May 2015.

13. 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.