An Analytical Study of Big Data Clustering Algorithms and its Challenges

P. Padmavathi¹, Dr. V.P. Eswaramurthy²
¹P.H.D Scholar, Periyar University, Salem-11, INDIA
²Assistant Professor, Government Arts College, Kumarapalayam, INDIA

ABSTRACT

Big Data is basically vast amount of data which cannot be effectively processed, captured, and analyzed by traditional database and search tools in reasonable amount of time. Big Data information explosion is mainly due to the vast amount of data generated by social media platform, data input from many channels, various mobile devices, user agents, multimedia data, and so on. Overall it is an expanding “Digital Universe”. Big Data predominantly revolve around 5V’s: Volume, Velocity, Variety, veracity, Value. Big data plays a major role in all business sectors in the digital era.

The purpose of this research is to analyze the challenges of bigdata , tools and techniques used and discuss different clustering methods including Data Mining clustering algorithms, dimension reduction techniques, parallel classification and the Map Reduce framework and its challenges.

Keywords-- Big Data, social media, Digital Universe, Volume, Velocity, Variety, veracity, Value, Map Reduce.

I. INTRODUCTION

Big Data is a large volume of data from various data sources such as social media, web, genomics, cameras, medical records, aerial sensory technologies and information sensing mobile devices. Big Data includes structured, semi-structured and unstructured data. This unstructured data contains useful information which can be mined. Since 1980s per capital capacity to store information is increased into double the amount for every 40 months. In 2012, statistics say that 2.5 quintillion bytes of data are created per day. Moreover, digital streams that individuals create are growing rapidly. For example, most of the people are using camera on their own. Big Data are of high level volume, high velocity, and high variety of information that needs advanced method to process the Big Data. In addition, conventional software tools are not capable of handling Big Data. So Big Data requires extensive architecture. The 5vs plays essential role in bigdata which includes Volume, Velocity, Variety, veracity and Value.

Volume refers to the vast amounts of data generated every second. Just think of all the emails, twitter messages, photos, video clips, sensor data etc. we produce and share every second. We are not talking Terabytes but Zettabytes or Brontobytes. On Facebook alone we send 10 billion messages per day, click the "like" button 4.5 billion times and upload 350 million new pictures each and every day. If we take all the data generated in the world between the beginning of time and 2008, the same amount of data will soon be generated every minute! This increasingly makes data sets too large to store and analyze using traditional database technology. With big data technology we can now store and use these data sets with the help of distributed systems, where parts of the data is stored in different locations and brought together by software.

Velocity refers to the speed at which new data is generated and the speed at which data moves around. Just think of social media messages going viral in seconds, the speed at which credit card transactions are checked for fraudulent activities, or the milliseconds it takes trading systems to analyse social media networks to pick up signals that trigger decisions to buy or sell shares. Big data technology allows us now to analyse the data while it is being generated, without ever putting it into databases.

Variety refers to the different types of data we can now use. In the past we focused on structured data that neatly fits into tables or relational databases, such as financial data (e.g. sales by product or region). In fact, 80% of the world’s data is now unstructured, and therefore can’t easily be put into tables (think of photos, video sequences or social media updates). With big data technology we can now harness differed types of data (structured and unstructured) including messages, social media conversations, photos, sensor data, video or voice...
analytics technology now allows us to work with these type of data. The volumes often make up for the lack of quality or accuracy.

Value: Then there is another V to take into account when looking at Big Data: Value! It is all well and good having access to big data but unless we can turn it into value it is useless. So you can safely argue that ‘value’ is the most important V of Big Data. It is important that businesses make a business case for any attempt to collect and leverage big data. It is so easy to fall into the buzz trap and embark on big data initiatives without a clear understanding of costs and benefits.

II. REVIEW OF PREVIOUS WORK

Zomaya et al. 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 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 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. Naggpal and Mann’s paper does not address all the clustering technique it is interested only to study density based clustering algorithms such as DBSCAN, DENSITYCLUE and to discuss their advantages and disadvantages. 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.

III. BIG DATA CHALLENGES

Big data sizes are constantly increasing, currently ranging from a few dozen terabytes (TB) to many petabytes (PB) of data in a single data set. Consequently, some of the difficulties related to big data include capture, storage, search, sharing, analytics, and visualizing. Today, enterprises are exploring large volumes of highly detailed data so as to discover facts they didn’t know before. Hence, big data analytics is where advanced analytic techniques are applied on big data sets. Analytics based on large data samples reveals and leverages business change. However, the larger the set of data, the more difficult it becomes to manage.

Indeed, the extreme challenge of Big Data is to make heterogeneous data (temperature, 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. According to HACE (Heterogeneous, Autonomous, Complexity, Evolving) theorem the most important characteristics of Big Data are:

Heterogeneous data, that’s means that 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. 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, 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). 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. 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. 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.
IV. BIG DATA ANALYTICS TOOLS AND METHODS

With the evolution of technology and the increased multitudes of data flowing in and out of organizations daily, there has become a need for faster and more efficient ways of analyzing such data. Having piles of data on hand is no longer enough to make efficient decisions at the right time. Such data sets can no longer be easily analyzed with traditional data management and analysis techniques and infrastructures. Therefore, there arises a need for new tools and methods specialized for big data analytics, as well as the required architectures for storing and managing such data. Accordingly, the emergence of big data has an effect on everything from the data itself and its collection, to the processing, to the final extracted decisions. Consequently, proposed the Big –Data, Analytics, and Decisions (B-DAD) framework which incorporates the big data analytics tools and methods into the decision making process. The framework maps the different big data storage, management, and processing tools, analytics tools and methods, and visualization and evaluation tools to the different phases of the decision making process. Hence, the changes associated with big data analytics are reflected in three main areas: big data storage and architecture, data and analytics processing, and, finally, the big data analyses which can be applied for knowledge discovery and informed decision making. Each area will be further discussed in this section. However, since big data is still evolving as an important field of research, and new findings and tools are constantly developing, this section is not exhaustive of all the possibilities, and focuses on providing a general idea, rather than a list of all potential opportunities and technologies.

V. CLUSTERING METHODS

1.5.1 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. 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.

1.5.2 Multiple-machine 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. It includes Parallel clustering and Map Reduce based clustering algorithms.

1.5.3 MapReduce

It is a programming model and an associated implementation for processing and generating large data sets with a parallel, distributed algorithm on a cluster. A Map Reduce program is composed of a Map() procedure (method) that performs filtering and sorting and a Reduce() method that performs a summary operation (such as counting the number of students in each queue, yielding name frequencies). The Map Reduce System (also called “infrastructure” or “framework”) orchestrates the processing by marshalling the distributed servers, running the various tasks in parallel, managing all communications and data transfers between the various parts of the system, and providing for redundancy and fault tolerance.

VI. 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 challenges raised by the Big Data. Indeed there is no clustering algorithm that can be used to solve all the Big Data issues and there are many possibilities to research on the same. Although the parallel classification is potentially very useful for Big Data clustering, the complexity of the implementation of these algorithms remains a great challenge which will need more computing power. However, the Map Reduce framework can support 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

[2] Drowning in numbers -- Digital data will flood the planet---and help us understand it


