Big Data: Characteristics, Issues and Clustering Techniques

Nandita Yambem[1*], and A.N. Nandakumar[2]

[1] VTU-RRC, Bangalore, India

[2] Department of CSE, GSSSIETW, Mysore, India

DOI: https://doi.org/10.21467/proceedings.1.55

* Corresponding author email: nanditayambem@gmail.com

Abstract

Now a days as Big Data is generated from sensor networks, IOTs, Social network and many collaborative networks, etc. continuously over a period of time line resulting in very large volumes of Data of various types at a phenomenal rate. Handling the characteristics of Big Data is a huge challenge. To handle this large data called big data a different approach or algorithm is required by organizations to handle all the challenges of Big data the Big data system requires a massive processing power and stable complex network configurations. Focuses should be on the analysis part of the big data classification by implementing different techniques in it. The various clustering techniques and the algorithms with the challenges they pose with Big data are discussed. To handle the dimensionality issues and approach the unsupervised machine learning, the use of clustering technique is the most common step in Big Data.

Index Terms- Big Data, Issues, Data - security, complexity, heterogeneity, clustering techniques etc.

1 INTRODUCTION

Big data refers to datasets with sizes beyond the ability of commonly used software tools to capture, accurate, manage, and process the data within a tolerable elapsed time which means data is too big, fast and too hard for existing systems of database management tools, traditional data processing applications and algorithms to handle [1][2][3]. The sizes of big data are constantly moving target, as the size ranging from a few dozen terabytes to many petabytes of data in a single data set. Big Data technology improves

i. performance,

ii. facilitates product innovation and business model's services,

iii. provides decision making support.

iv. minimize hardware and processing costs

v. the value of Big Data prior to committing significant company resources.

Big data sources can be from

i. Web data, e-commerce
ii. purchases at department/grocery stores
iii. Bank/Credit Card transactions
iv. Social Network

Hence, Big Data applications can be applied in various complex scientific disciplines, single or interdisciplinary, including atmospheric science, astronomy, medicine, biology, genomics, and biogeochemistry.

1.1 CHARACTERISTICS OF BIG DATA

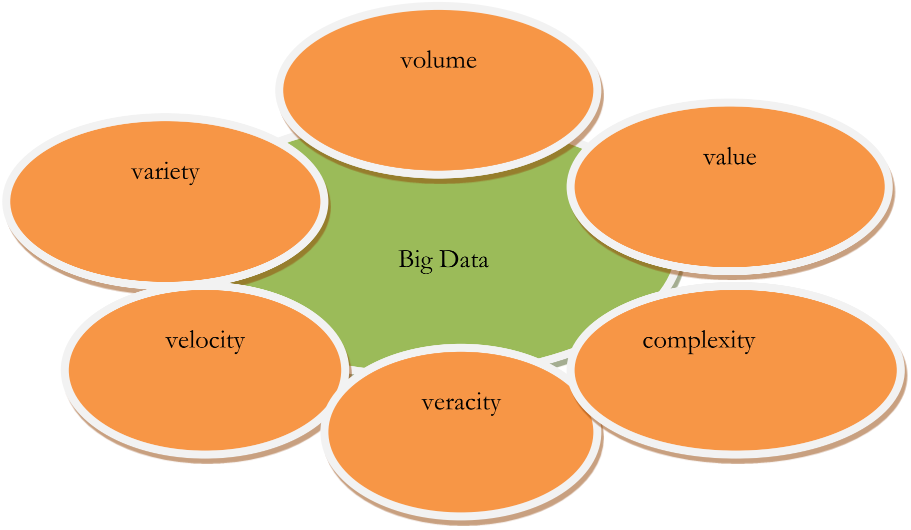

Fig 1: The 6V's of Big Data

1.1.1 Volume

Clustering algorithms have the ability to deal with large amount of data. Volumes measure the amount of data available to an organization. As volume increases , the value of data record decreases in proportion to age, type , richness and quantity among other factors[1][2][3].

The Volume property criteria for clustering algorithms considered are

i. Size of the data set
ii. High dimensionality
iii. Handling Outliers

Size of the data set refers to the collection of attributes which are categorical, nominal, ordinal, interval and ratio. Many of clustering algorithms support numerical and categorical data. High dimensionality is to handle big data as the size of data set increases number of dimensions also increases. Outliers: Many clustering algorithms are capable of handling outliers and noise data.

1.1.2 Variety

The ability of a clustering algorithm to handle different types of data (numerical, categorical and hierarchical) from different sources whose major challenge is storing, retrieving these data types quickly and cost efficiently and aligning data types from different sources to describe a single event to extract and analyse.

While selecting a clustering algorithm with respect to the Variety property, consideration is on

i. type of dataset i.e. size of data set is small/big. Many of the clustering algorithms supports large data sets for big data mining and

ii. clusters shape i.e. depends on the data set size and type shape of the cluster formed.

The varied data are contained in websites, blogs, emails, exchanges on social networks (Facebook, Twitter, LinkedIn ...), images, video, audio, logs, data spatial (geolocation), the biometrics, etc.[1][2][3].

1.1.3 Velocity

Data is flooding at a very high speed and needs to be handled within reasonable time. One of the challenges in big data is responding quickly to data velocity . Velocity describes data in motion i.e.speed at which data is generated/created, streaming and aggregated. The advent of e-commerce has rapidly increased the speed and richness of data used for different business transactions (for example, web-site clicks) [1][2][3][4].

The computations of clustering algorithm based on velocity is

i. running time complexity of a clustering algorithm.

ii. run time performance

e.g.: Real Time Big Data Analytics

For example, when you visit a sophisticated content web site such as Yahoo, the ads that pop up have been selected specifically based on the capture, storage, and analysis of your current web visit, your prior web site visits, and a mash up of external data stored in a NoSQL DB like Hadoop and added to the analytics.

1.1.4 Veracity

It is the biases, noise and abnormality in data. It is the uncertainty due to data inconsistency and incompleteness, ambiguities, latency, deception, model approximations. Big data is sourced from many different places, hence it is required to test the veracity/quality of the data [3][4][5].

1.1.5 Value

Input parameters plays an important role for a clustering algorithm to process the data accurately and form a cluster with less computation. Hence, value deals with the usefulness of data in making decisions [3][5][8].

1.1.6 Complexity

Data from different sources have different structures and needs to be connected and correlate relationships and linkages.

Complexity measures of the degree of interconnectedness and interdependence in big data structures such that a small change or combination of small changes in one or a few elements can yield very large changes or a small change that ripple through the system ,substantially affecting its behaviour, or no change at all[2][6]. It is difficult to process and analyse complex data because of relationship between different variables in the same dataset or multiple datasets. Traditional analysis tools, takes multiple iterations to understand relations between variables. To discover relationship between variables, complex data analysis can take advantage of distributed computing of big data [8][9][10].

2 big data issues

The Big Data fundamental issue areas that needs to be addressed are [9][10][11][12][13][14][15][16]

i. Data Challenges
ii. Process Challenge
iii. Management Challenge
iv. Transmission and Storage

2.1 Data Challenges

Data quality:

Data needs to be accurate or timely, otherwise the value for decision making will be jeopardized and more complicated considering the volume of information in big data projects.

Data speed:

The challenge in going through volumes of data and accessing level of details needed at high speed grows as the degree of granularity increases.

Data Discovery:

Identifying the right data from the vast amount of data is a big challenge. As large number of sources such as social networking sites, blogs, different types of content such as articles, comments, companies find it difficult to identify the right data and determine how to make the best use of it, there is the need to find out the rules that will help in identifying the right data.

Huge challenge is on how to find high-quality data from the vast collections of data that are available on the Web.

Data Relevance and Comprehensiveness

A lot of understanding needed to get data in the right shape so as to be visualized as part of data analysis. for example:-if the data is from social media content, than, information of the user – like customer using a particular set of products needs to be known– and understand what it is trying to visualize out of the data. Visualization tools fail without some sort of context. Hence, proper domain expertise is required to deal such challenge. People analysing the data should have deep understanding of the source of data, who will consume the data and how the information will be interpreted by the user.

Data Scalability:

It is a major challenge in big data and needs effective solution to enable cost effective, scalable storage and processing of large volume of data. Most NoSQL solutions like Mongo DB or HBase have their own scaling limitations.

Data security and privacy:

Security is the big concern with the big data. As larger amount of data is processed and transferred among the organizational boundaries, the Big data carries greater risks if it contains credit card data, personal ID information and other sensitive assets. The challenge is how to protect this sensitive data keep it private. To safeguard Big Data, most NoSQL big data platforms have few security mechanisms. Big data sizes ranges from a few terabytes to many petabytes of data in a single data set. As big data expands, each data should be verified and techniques to identify malicious data. For this, analysis of massive amount of data will be correlated, analysed and mined for meaningful patterns. To meet the Security control of Big Data following are the requirements:

i. Basic functionality of the cluster must not compromise.
ii. Should be scale as the cluster.
iii. Essential big data characteristics must not compromised
iv. Big data environments security threats to be addressed or data to be stored within the cluster.

Three categories of security violation are unauthorized release, information modification and denial of resources. Some of the security threats:

i. Accessing of files and execute arbitrary code and carry out further attacks by unauthorized user.
ii. Unauthorized user eavesdrop/sniff to data packets being sent to client
iii. Unauthorized client performs read/write on a data block of a file
iv. May gain access privileges and may submit a job to a queue or delete or change priority of the job by unauthorized client.

Some methods used for protecting big data are:

i. Authentication methods like Kerberos can be used for verifying user or system density.
ii. File encryption methods ensuring user information confidentiality and privacy and secures the sensitive data. Regardless of OS/platform type ,the file layer encryption provides a consistent protection across different platforms.
iii. Specifying access control privileges to enhance security for user or system.

iv. Key management services to distribute keys and certificates and manage different keys for each group, application, and user.

v. Logging to manage log files and to look when something fails or hacked. For security requirements the entire system needs to be audited on a periodic basis.

vi. Use secure communication between nodes and applications requiring an SSL/TLS implementation thus protecting all network communications

Two common approaches are used to protect privacy:

i. Adding certification or access control to the data thus restricting access to data. The challenge is to design a secured certification/access control mechanisms.

ii. Data fields to be anonymize to avoid pinpointing sensitive information to an individual record

Data representation:

For computer analysis and user interpretation data needs to be more meaningful. If data representation is improper it will reduce the value of the original data and may obstruct effective data analysis. To enable efficient operations on different datasets, efficient data representation shall reflect data structure, class, and type, as well as integrated technologies.

Data access and sharing:

Though there is a potential value for development, most of the publicly available online data, much more valuable data are held by corporations and is not accessible. There is reluctance of private companies and other institutions to share data about their clients and users, as well as about their own operations. When data is stored in places and difficult to be accessed, transferred, etc., it leads to institutional and technical challenges.

Data Heterogeneity and Incompleteness:

Heterogeneous data comes from several different sources like Twitter, Facebook, LinkedIn and instant messaging are in complex and heterogeneous format requiring a set of techniques and the implementation of various solutions. Major challenge is to figure out what is the data one has and how to analyse it which requires adapting and integrating multiple analytic. Data can be both structured and unstructured. Prior to data analysis data must be carefully structured.

Data are highly dynamic and does not have particular format. For example email attachments, images, , X rays, voice mails, graphics, pdf documents, medical records ,video, audio etc. and they cannot be stored in row/ column format as structured data. Incomplete data creates uncertainties for data analysis and must be managed during data analysis. Such values are caused by different realities, like sensor node malfunction, or some systematic policies to intentionally skip some values.

Displaying meaningful results

When dealing with extremely large amounts of information or a variety of categories of information plotting points on a graph for analysis becomes difficult.

One solution is to cluster data into a higher-level view where smaller groups of data become visible i.e. data can be effectively visualized by grouping the data together.

Dealing with outliers

The graphical representations of data made possible through visualization communicates trends and outliers much faster than tables containing numbers and text. Users can easily spot issues needing attention by just glancing at a chart.

Outliers typically represent about 1 to 5 percent of data but viewing 1 to 5 percent of the data is rather difficult while working with massive amounts of data.

Data Complexity

The data is collected from various contexts (multi-source, multi-view, multi-tables, sequential, etc.) as well as from decentralized treatment data or massively parallel processing (MapReduce) With the increase in volume data complexity increases and the usual treatment methods, with management of relational database tools are not sufficient enough to meet the requirements capture, storage and further analysis.

2.2 Process Challenge

Process challenges include:

i. Capturing data, wherein critical piece of information is discovered and extracted providing with leverage in some situation

ii. Data aligning from different sources.

iii. Transforming the data suitable for analysis, so as to select the best, but not necessarily the optimal solution.

iv. Modelling it, whether mathematically, or through some form of simulation

v. Understanding the output, visualizing it and sharing the results.

2.3 Management Challenge

Many data warehouses have sensitive data like personal data which are of legal and ethical concerns for accessing such type of data. Hence, data security and access controlled as well as logged for audits must be maintained.

The main management challenges are

i. Data privacy

ii. Security

iii. Governance

iv. Ethical

2.4 Transmission and Storage

The quantity of data explodes each time a new storage medium is invented.

i. Everyone and everything is creating data by (e.g., devices, etc.) – by professionals like writers, scientist, journalists etc. Current disk technology limits are about 4terabytes per disk i.e. 1 exabyte would require 25,000 disks. It would be unable to directly attach the requisite number of disks, even if an exabyte of data could be processed on a single computer system. The current communication networks data would be overwhelmed on access to that data.

ii. Assuming that a 1 gigabyte per second network has an effective sustainable transfer rate of 80%, the sustainable bandwidth is about 100 megabytes. Thus, if a sustained transfer could be maintained, transferring an exabyte would take about 2800 hours. To transmit the data from a collection or storage point to a processing point it will take longer than to actually process it.

iii. Two solutions manifested are-

   a. First, the data in place is processed and only the resulting information transmited, i.e. the code to the data is brought instead of bringing the data to the code.

   b. Second, perform triage or ordering on the data and transmit only critical data to downstream analysis maintaining integrity and the provenance metadata which should be transmitted with the actual data.

3 Big Data clustering technique

Big Data clustering techniques can be classified into two categories [17][18] :

i. single machine clustering techniques

   a. Sample Based (K-means, CLARANS, BIRCH, DBSCAN, STING, EM, CURE)

   b. Dimension Reduction (Locality-Preserving Projection, Global Projection)

ii. multiple machine clustering techniques

   a. Parallel clustering (Parallel K-means, Parallel Fuzzy c-means, DBRC, ParMETIS)

   b. MapReduce clustering (MR- DBSCAN, MapReduce Based on GPU)

3.1 Single-machine clustering techniques

Single-machine clustering algorithms run in one machine and can use resources of just one single machine.

3.1.1 Sample based technique

Clustering algorithms is performed on a sample of dataset and is generalized to the whole dataset instead of clustering the whole dataset .

i. Deals with exponential search space.

ii. Improves speed and scalability

The challenges involved are - poor handling of noisy data and outliers, works only on numeric data, random initial cluster centre problem , empty cluster generation problem, order-sensitive and may generate different clusters for different orders of the same input data , may not work well for non-spherical clusters , not suitable for high-dimensional datasets , quality depends upon the threshold set.

3.1.2 Dimension reduction technique:

Dimensionality of dataset is an important aspect as more the dimensions the data have, the more is the complexity and longer execution time. Sampling techniques reduce the dataset size but doesn't support for high dimensional datasets.

Size of data are measured in two dimensions

i. Number of variables

ii. Number of examples

During exploration and analysis of these data, these dimensions can take very high values which may cause a problem. So before applying clustering algorithm it is essential to implement data processing tools and make a pre-treatment to the dataset.

The Dimension reduction technique selects or extracts optimal subset of relevant features for criteria already fixed eliminating irrelevant and redundant information based on the criterion making it possible to reduce the size of the sample space and more representative of the problem. To avoid the disadvantages of high dimensionality for large sets of data dimension reduction is usually performed before applying the classification algorithm.

Feature selection process is to select from a set of original variables, an optimal subset of variables, according to a certain performance criteria to reduce the number of required actions, than a parallel k-means algorithm is applied to the data subsets. Feature selection provides better classification accuracy and takes much less time than existing algorithms other classification algorithms for Big Data.

Feature extraction

i. select features in a transformed space - in a projection space

ii. use all the information to compress and produce a vector of smaller dimension.

The challenges involved are - no efficient solution for high dimensional datasets hence it should be performed before applying the classification algorithm.

3.2 Multiple –machine clustering techniques

Multiple-machine clustering techniques runs in several machines and has access to more resources. The scalability and speed of the algorithms improves the sampling and dimension reduction methods, but the growth of data size is much faster than memory and processor advancements, consequently one machine with a single processor and a memory cannot handle terabytes and petabytes of data thereby needing algorithms that can be run on multiple machines.

This technique breaks the huge amount of data into smaller pieces which can be loaded on different machines and the processing power of these machines is used to solve the huge problem. Multi machine clustering algorithms are divided into two main categories:

i. Un-automated distributing– parallel

a. Parallel clustering is very complicated and time consuming because of developers involved with parallel clustering challenges, and data distribution process details between different machines available in the network as well.

ii. Automated distributing– MapReduce

a. MapReduce help programmers from unnecessary networking problems and concepts such as load balancing, data distribution, fault tolerance etc. by automatically handling them allowing huge parallelism, easier and faster scalability of the parallel system.

The challenges involved are - complexity of implementing the algorithms difficult, implementing each query as a MR program is difficult, and no primitives with respect to common operations (selection/extraction).

4 CONCLUSIONS

This technical paper covers key issues and characteristics related to Big Data. We have given brief insights into 5 V's (Volume, Velocity, Variety, Varacity, Value) and 1 C(complexity), Big Data key issues like Data, Process, Management, Transmission and Storage Challenges have been elaborated at micro levels . Basic Big Data Clustering Techniques have been explored of which MapReduce is of paramount importance to my ongoing research on Big Data Analytics.

References

[1] K.Arun, Dr.L.Jabasheela, Big Data: Review, Classification and Analysis Survey, IJIRIS, Volume 1 Issue 3 (September 2014)

[2] Nawsher Khan, Ibrar Yaqoob, Ibrahim Abaker Targio Hashem, Zakira Inayat, Waleed Kamaleldin Mahmoud Ali, Muhammad Alam, Muhammad Shiraz, and Abdullah Gani Big Data: Survey, Technologies, Opportunities, and Challenges, The Scientific World Journal Volume 2014 (2014)

[3] C.L. Philip Chen , Chun-Yang ZhangData-intensive applications, challenges, techniques and technologies: A survey on Big Data, 2014 Elsevier

[4] Harshali H. Deshmukh,Big Data Analytics, Transactions on Engineering and Sciences - TECH-KNOW DOCX – 2015

[5] Stephen Kaisler ,Frank Armour,J. Alberto Espinosa,William Money,Big Data: Issues and Challenges Moving Forward, 2013 46th Hawaii International Conference on System Sciences

[6] Alexandru Adrian TOLE ,Big Data Challenges , Database Systems Journal vol. IV, no. 3/2013

[7] Ali Seyed Shirkhorshidi, Sr Aghabozorgi, Teh Ying Wah, Tutut Herawan,Big Data Clustering: A Review, Conference Paper · June 2014, https://www.researchgate.net/publication/267395205

[8] Ms. Kirti P. Sahare, S. A. Murab, Dr. M. V. Sarode, M. M. Ghonge, BIG DATA: The Leading Innovative and Productive Framework, IJSRET, Volume 2 Issue 12 pp 891-896 March 2014

[9] Bharti kalra, Suryakant Yadav, Dr. D.K. Chauhan ,Review of Issues and Challenges with Big Data, International Journal of Computer Science and Information Technology Research, Vol. 2, Issue 4, pp: (97-101), Month: October - December 2014

[10] Piyush Gupta1, Pardeep Kumar Mittal2, Girdhar Gopal3 Big Data: Problems, Challenges and Techniques, JECET; March 2015-May 2015; Sec. B; Vol.4.No.2, 202-209.

[11] Roberto V. Zicari , Big Data: Challenges and Opportunities, ODBMS.org www.odbms.org

[12] Min Chen · Shiwen Mao · Yunhao Liu, Big Data: A Survey , Springer Science+Business Media New York 2014

[13] Jaseena K.U and Julie M. David, ISSUES, CHALLENGES, AND SOLUTIONS:BIG DATA MINING , NeTCoM, CSIT, GRAPH-HOC, SPTM – 2014 pp. 131–140, 2014. © CS & IT-CSCP 2014

[14] M.H.Padgavankar, Dr.S.R.Gupta , Big Data Storage and Challenges, M.H.Padgavankar et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014, 2218-2223

[15] Monica Bulger, Greg Taylor, Ralph Schroeder , Data-Driven Business Models: Challenges and Opportunities of Big Data , Oxford Internet Institute September 2014

[16] CHANGQING JI , YU LI , WENMING QIU , YINGWEI JIN, YUJIE XU , UCHECHUKWU AWADA , KEQIU LI , BIG DATA PROCESSING: BIG CHALLENGES AND OPPORTUNITIES , Journal of Interconnection NetworksVol. 13, Nos. 3 & 4 (2012) World Scientific Publishing Company.

[17] Min Chen, Simone A. Ludwig, and Keqin Li,Clustering in Big Data , "K29224_C016" — 2017/1/12

[18] S.Mahalakshmi, C.SaiAshwini , Meghana S , Research Study of Big Data Clustering Techniques , IJIRSE/Vol 4. Iss. 5/ Page 80