Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 6.017

IJCSMC, Vol. 5, Issue. 11, November 2016, pg.110 – 117

Wipe to Big Data Mining Applications and Several Challenges

Yasir Ali Mmutni Alanbaky

Dept. of computer science, basic education college, Diyala University, IRAQ, ee22a12@gmail.com

Keywords

Datasets,

Big Data,

Data mining,

Independent sources.

Abstract

Today to discover datasets a modern term appear "Big Data" because of large size of data and their complexity. It is hard to deal with these large amounts of data with the present techniques. A large volume of data in Big data, hard to understand it or analyze with multiple independent sources and growing in data sets size. The term Big Data mining is the ability of conclude useful information out of those great dataset, because of its big volume, Fastness, and differentiability it wasn't practical to do it previously. My study paper view totals the features and characteristics of "Big data", and challenges in "Big data" and its linked works. Additionally I focused in several article studies which written by most talented scientists on the field of "Big data mining". I trust my study will helpful to remodeling the current technologies of "data mining" to resolve the challenges "Big data".

1. Introduction

We can be sure that we have a big ability to collect data from various resources in different forms independent or linked application. Today's, the data quantity is formed every one day is predestined to be 2.5 Exabyte's . That big amounts of data is a new challenges give us discovering new tasks. The big rise amount of data is over our capability to processing it, analyze, recover, understand and store those datasets. The real time analytics to data stream need to treat the data generated recently , from several applications like :Facebook , blogging ,email, log records ,Twitter sensor networks etc [1].

Let's look to data of Internet . Web pages generated by Google were about one million in 1999, when it reaches to one billion in 2001 and it is more increased one trillion in 2008. Every and each two day entire world generate 5 quintillion bytes of data. Those is a lot of data , over and above the 90% of these data in the entire world now was generated in the past two years only [1]. That fast increase in data amount is hurried by the active increased in our social network applications, like face book , Weibo, Twitter, etc , these is giving permit to user to generating several contents freely and update the present large Web size data. after that ,with various mobile phone become a simple way to obtain the real time data, the so much volume of data that mobiles can process it to changing our regular life have safely rise our bygone calls data records founded processes for billing objective . We can see the different applications of Internet shall growing the volume of data to the new standard. The different personal devices such (buses, vehicles, airports and railway stations) and people are all loosely linked . Trillions of such joint elements generate a big data, and substantial informations have to discovering from these data to assist to modify the form of our life . with all applications I reported, we be faced with the problems of system capacity and how to solving the problem linked with that to several businesses model . So the "Big Data mining" is the answer.

2. Big Data Mining

"Big data" is primarily established out due to the reality that we generating a great amount of data every and each day. "Big Data mining" was so pertinent from the first , like the initial work mention "Big Data" was a "data mining" book also that appeared in " Weiss and Indrukya in 1998"[2] . Big Data is data only , which is available in independent , and heterogeneous sources, in quite large quantities , and which getting updating in fragments of second. We should believed that the Big data will played a role in the next days in our lives in all fields . Like ,the server of Facebook that stores different type of data , like most of us, every day we use Facebook, uploading different type of data, such videos ,and photos. All that data gets stored in the server of Facebook at the data warehouses. This big amount of data called the "big data". There is another examples are storing of photos, videos and different type of data at different websites. Those are a good examples of the "Big Data" at the time. A fresh sources of large data are appear which created from big companies as Facebook, Google, Yahoo, Twitter , Apple, and from using mobile devices , which start to looking closely to those data to discover a useful style to enhancing skills of users. The traditional Database management systems have no suitable abilities to deal with such big data , that is why we need new tools, and new algorithms to dealing with such big amount of data. The first one is talk on "3 V's Big Data management" was Doug Laney [3] . The (Velocity, Variety, and Volume) of big data are the 3 V's, each Represents one aspect of captious

insufficiency of today's "Database management systems". Those 3 V's (Velocity, Variety, and Volume) are dimensions of data or defining properties. The 3 V's has now become common frameworks to dealing with and understanding the "big data". Presently there are two V's come out , value and variability. The "value" showing business value which give organizations advantages that undeniable . The "variability" showing the data structure changes .

The three V's that defining "big data" are Velocity, Variety, and Volume .

i) **(Velocity):** Velocity can define by the speed that the data is being created. Different applications have various hiding requirements, when the "decision makers" need the information that is very necessary in too short time as possible . These quickness to generate data is referred to "Velocity" in "Big Data".

ii) **(Variety):** the term Variety refers to the different forms in which the data is starting generated and stored. Various applications generating and storing data in different forms . Today , there is massive amount of "unstructured data" start generating aloof from "structured data" which generating from Business. Even the improvement in the technologies of "Big Data". the manufacture did not have any reliable and powerful (technologies and tools) which can working with a huge unstructured data as see now. Todays, the organizations not need only to depend on the "structured data" from databases, the organizations also compelled to consuming the data which is start generating together outside and inside the enterprises such that social media, clickstream data, etc. to remain competitive.

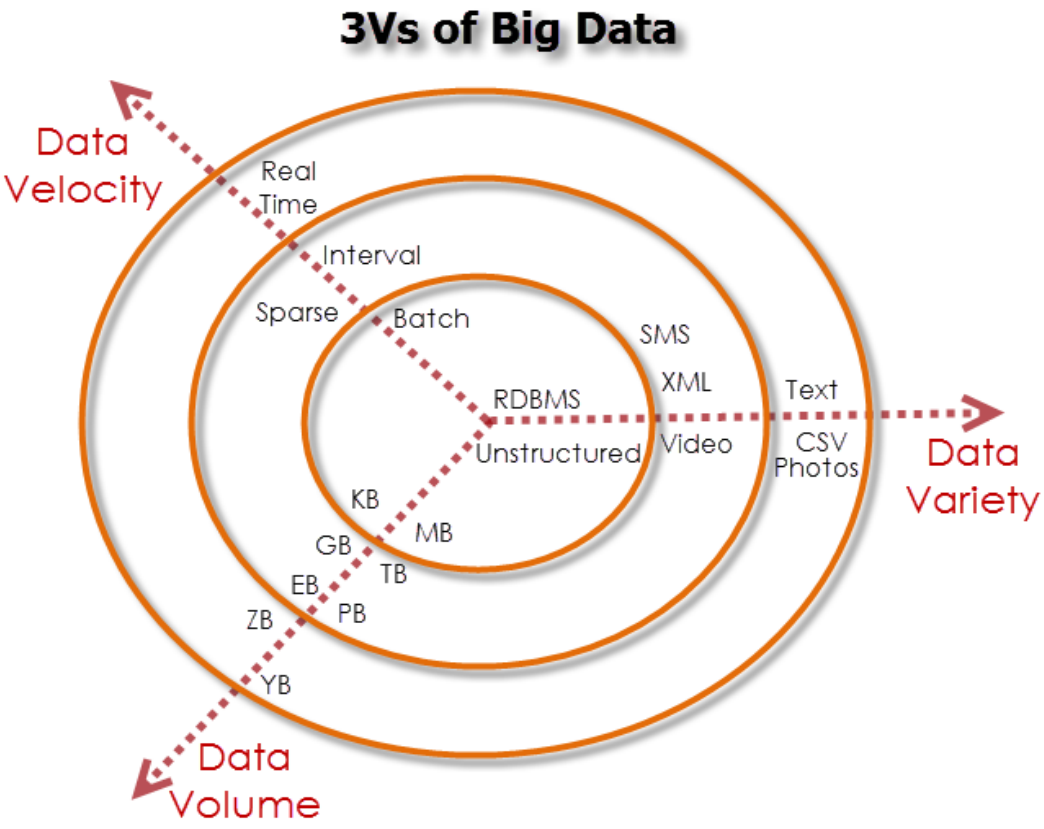

Figure 1: The 3 V's of "Big Data"

iii) **(Volume):** It is the size of data that we dealing with it. Through the advantage of technologies and with the discovery of various programs of social media, the data quantity is rising too quickly. Those data is spreading through the various area, in different forms , in great amounts range from Gbytes to Tbytes, Pbytes, and more and more. Todays, the data is not generating by people only, also a major amounts of data is start generating by a machine and it is exceed that generated by human. These side of data size is called "Volume" in "Big Data".

A) Different fields of big data applications

I am will explain in this paper the big data how is used today and with adding a real value. All side of our lifetime will be affected via "big data". Categorizing the applications of big data is too important to several category when we seeing what is the more used and highest benefits.

i) Optimize Public Health and Healthcare

The power of computing analytic the big data let us to decoding whole "DNA" series in moments , which will allowing researchers to better understand and finding new cures . Think about it what happened while calculating all personal health information
from wearable devices and smart watches could be using to applying it for millions of various people's diseases. Already, the techniques of big data used to observe child in the sick baby unit and specialist premature . In this way, the medical team would intervening in the time to saving premature babies in all time.

i) Optimize big business Processes

The Big data is increasing in business processes. Increased traders stock depend on prognosis created by web search trends, weather forecasts , and social media information's . A special process of business which have a many analysis steps of big data is delivery route optimization or supply chain. To optimize business Processes , by using sensors of "radio frequency identification" and "geographic positioning" to tracking delivery vehicles or goods and optimizing routes by supplying the information about live traffic.

iii) Optimize Performance of Device and Machine

Analytics of big data are helpful in optimize performance to Device and Machine. Such as, the tools of big data are utilized to run self-driving vehicle by Google. Like the cars of Toyota Prius is fitted with GPS , cameras, sensors and computers to driving safely on the roads without the human beings intervention. The tools of big data are used also to optimizing grids of energy by using smart meters data. The tools of big data are also used to optimizing the data warehouses and performance of computers.

iv) Optimizing and Improving Countries and Cities

The "Big data" using to improving several sides of our countries and cities . Like , big data allowing our cities to optimizing flows of traffic and improving information of real time traffic as well as and weather data and social media. There are many cities are trying to analytics of big data to turn it to "Smart Cities ", where it joining the transfer utility processes and infrastructure together.

B) Big Data Challenges

i) Mishandling in mining big data

In here , the challenges with one another with possible big data misuse are the issues, since the power in information. In future , unknown types of the data that produced by people also an issue. To cope those challenges we should increase and strengthen our capacity and our intent [5].

ii) Energetic Purveyance

Providing the service of cloud computing ,where it is the infrastructure such services which it are active resources if it required, much cloud computing linked communities are implement these idea also to make it simple to access to these services for clients . The new frameworks don't have energetic Purveyance property. Cases are in here that computing resource could become imperfect for the offer jobs, somewhat processes may be required additional resources. Other case , the protection and scheduling algorithms, existing algorithms don't considering those sides [4] .

iii) Privacy and Security of Big Data

Protection of cloud computing alliance with working group of big data identified good protection and isolation problem which we needs to reserve to making the infrastructure and big data computing more safe. Generality of those cases are related to the computation and storage of big data. We have a few challenges linked to securing storage of data [6]. Various challenges of security linked to data privacy and data security are debated in [7] since it including data reliability, data breaches, data support, and data accessibility .

iv) Designing algorithms of Big data mining

At different locations the Big Data is stored , likewise the volumes of data are getting expand like the it keeps increasing constantly. It is too costly because gathering all data stored at different places. Let's think about it, if the classic methods of data mining are used for mining of Big Data, which are using to mining a small amount of data in computers system , then it will be difficult for it. One of the main goals of data mining algorithms is to protect the privacy of the data. Separated the large data sets to number of subsets , then the mining algorithms will applied to these subsets. After that , will applied summation algorithms to resulting the mining algorithms , to get the aim of big data mining. In such method we will breaking the privacy statement while splitting one big data to smaller dataset. We will facing another challenges when we design such algorithms. Consider that , big data is set of bulky

and complex of data sets which it hard to processing it , and mine for knowledge and patterns used classic database management tools.

v) Building a unifying global system to mining in big data

There are much methods which are planned for carrying out grouping or classification independently, but no theoretical background there, that unifies different responsibilities like grouping, association guidelines , classification and so on. Therefore , constructing a global unifying model of big data mining is a dynamic field to research.

C) Related work

Big data have a large, dynamic , and dissimilar features of applications data engaged in a circulate environments , therefore this Big Data have to doing computing on the (PB) petabyte , even with (EB) Exabyte-level data but with complicated processes of computing. So, utilize computing equivalent to infrastructure, its conform software models to proficiently analyze and conform to support programming language , and mining the spreaded data are the main aims for Big Data processing to changing it the "extent" to "excellence". presently in the mining platform section, being used equivalent programming models such as "MapReduce" for mining of data and also for the perseverance of study, furthermore for the public there have a cloud computing platforms of Big Data services. "MapReduce" is a programming model and an associated implementation for processing and generating large data sets with a parallel, distributed algorithm on a cluster. There is yet a particular "gap" in performances with relational database. Accept a big volume of attention by improve the real-time nature of large-scale data processing with performance of "MapReduce", it being applied to data mining algorithms and many of machine learning. The algorithms which used for data mining mostly need to scanning out of data in the training for profit the statistics to improve or explain models.

We can found the significant works about big data mining in the international journals or main conference such as ECMLPKDD, ICDM, KDD. A Dmitriy Ryaboy and Jimmy Lin present a scaling Big Data mining infrastructure. A Jiawei Han and Yizhou sun present one paper in which they shows that mining heterogeneous information networks is a new and auspicious research leading edge in Big data mining. My study showing new cases of tools of data mining, it is not simple to performing analytics. It is presents experience of doing analytics at Twitter, and it is insight of Big data mining infrastructure.

U kang and christors faloutstos (A Big graph mining: Algorithms and discoveries) gives an summary of big graphs mining, focusing on using tool of the "PEGASUS" , that displaying proven results in Twitter social network and web graph . These paper showing promising future to exploration ways for big graph mining.

A few of people, whom expecting to employ another party like accountant to doing their data, so it is so important to have operative and effective contact to the data. Like these case, the user privacy restriction may be faced such no copying allowed or limited copies, etc. Therefore, there is planned "privacy-conserving" auditing mechanism for free for data storage on large scale. In [8] the popular "key based mechanism" is using to allowing "third party auditing", then users could permitting a third party securely to evaluating their data without give up the data privacy or breaking the security settings. In state of designing algorithms of data mining, the progress of Knowledge in actual world systems is a public occurrence. However as the changing of problematic statement, the knowledge will change consequently. For that [9], [10] and [11] started and suggested the notion of "local pattern analysis", which was put the groundwork for knowledge in worldwide invention in data mining from multi source. These notion delivered a solution not for the problems of complete examine only , also it delivered a view for discovering prototypes for all worldwide that ancient methods data mining that cannot be find.

D) Conclusion

In this study we studied the Big data concept in "data mining". The "Big Data" will have very big growing data during the upcoming years, therefore any data management developer have to managing so more amount of data at each years. These data is will be more quicker, bigger, miscellaneous. About the subject we discussed some approaches and the major challenges for the future which are the major concerns. Currently, Big Data mining helping us to discovering new knowledge. The rising in every engineering and science domains make us in need to Big data mining. We hope with Big data technologies will have the ability to providing most accurate and most relevant feedback of social sensing to best understanding at real time of our society.

References

[1] J. Gama. Knowledge discovery from data streams .Chapman & Hall/CRC, 2010.

[2] S. M. Weiss and N. Indurkhya. Predictive data mining practical guide. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1998.

[3] D. Laney. 3-D Data Management: Controlling Data Volume, Velocity and Variety. META Group Research Note, February 6, 2001.

[4] M. N. Vijayaraj, M. D. Rajalakshmi, and M. C. Sanoj, "Issues and challenges of scheduling and protection algorithms for proficient parallel data processing in cloud."

[5] U. G. Pulse, "Big Data for development: challenges & opportunities," NacionesUnidas, Nueva York, mayo, 2012.

[6] “Top ten big data security and privacy challenges,” Cloud Security Alliance White paper, 2012.

[7] A. A. Soofi, M. I. Khan, R. Talib, and U. Sarwar, "Security Issues in SaaS Delivery Model of Cloud Computing," 2014.

[8] C. Wang, S.S.M. Chow, Q. Wang, K. Ren, and W. Lou, "Privacy- Preserving Public Auditing for Secure Cloud Storage" IEEE Trans. Computers, vol. 62, no. 2, pp. 362-375, Feb. 2013.

[9] X. Wu and S. Zhang, "Synthesizing High-Frequency Rules from Different Data Sources," IEEE Trans.Knowledge and Data Eng., vol. 15, no. 2, pp. 353-367, Mar./Apr. 2003.

[10] X. Wu, C. Zhang, and S. Zhang, "Database Classification for Multi-Database Mining," Information Systems,vol. 30, no. 1, pp. 71- 88, 2005.

[11] K. Su, H. Huang, X. Wu, and S. Zhang, "A Logical Framework for Identifying Quality Knowledge from Different Data Sources," Decision Support Systems, vol. 42, no. 3, pp. 1673-1683, 2006.

