IJCAT - International Journal of Computing and Technology, Volume 3, Issue 4, April 2016
ISSN : 2348 - 6090
www.IJCAT.org

A Study of Big Data Challenges and Opportunities

[1] Amarbir Singh, [2] Sarabjit Singh

[1] Department of Computer Science, Guru Nanak Dev University
Amritsar, Punjab 143001, India

[2] Department of Computer Science, Guru Nanak Dev University College Verka
Amritsar, Punjab 143001, India

Abstract - The Big Data" as a term has been among the biggest trends of the last three years, leading to an upsurge of research, as well as industry and government applications. As per the user demand and growth trends of large free data, the storage solutions are now becoming challengeable to protect, store and retrieve data. The days are not so far when the storage companies and organizations will start saying 'no' to store your valuable data or they will start charging a huge amount for its storage and protection. The flood of big data will lead to the zettabyte per year range with in a little time period. The major attributes of big data to be emphasized are volume, velocity, variety and veracity and it always looks like that the storage issue will be resolved in near future but it is a long duration challenge. In this paper we have analyzed the growth trend of big data and its future projection. We have also focused on the impact of the unstructured data on various concerns and we have also suggested some possible remedies to streamline big data.

Keywords **- Structured Data, Unstructured Data, Veracity, Hadoop.**

1. Introduction

Everyone is talking about big data, and it is believed that science, business, industry, government, society, etc. will undergo a thorough change with the influence of big data. Technically speaking, the process of handling big data encompasses collection, storage, transportation and exploitation. "Big Data" originally meant for the high volume of data that could not be processed efficiently using simple database methods and tools. As the data is increasing rapidly, each time a new storage concept was introduced in order to retrieve the data easily. Big data is basically a growth of data that can be both structured and unstructured in form. The standard expansion of Big Data has been emphasized on structured data, but most practitioners have come to know that most of the useful information resides in massive, unstructured information in the form of images, texts etc. Big Data is that amount of data that is just beyond our immediate grasp. We have to really work hard to deal with such pool of data [1]. It is an exponential growth of data and it's readily available to users when it's required. Big data is important to all the users either it is an Internet User, researchers or an organization customer. Big data is generally expressed as the four Vs: volume, velocity, variety and Veracity.

1.1 Volume

The volume of big data is the growth rate of database means how much data. There are many issues involved in the increase of data daily. Some data is valuable and genuine but the capacity of invaluable data what we said unstructured data is thousand times more. The daily uploading of Tera to Peta bytes unstructured data is an issue of worry.

1.2 Velocity

It is all about the access rate of data which generally deal with time and speed means how fast data can be processed. The real challenge is that how efficiently we are retrieving the data without loss and without any interruption. So to provide this V to the customer and users is the big challenges for organizations.

1.3 Variety

There is variety of data uploading and processing on Internet as well at local level of an organization. It is broadly divided in two categories structured and unstructured data. Data is available in variety of formats

like signal, recording, file and in content format. Challenge is to manage data in different formats.

1.4 Veracity

It suggests that despite the data being available, the quality of data is still a major concern. That is, we cannot assume that with big data comes higher quality. In fact, with size comes quality issues, which needs to be either tackled at the data pre-processing stage or by the learning algorithm.

2. Big Data Categories

There are millions of bytes data uploading and downloading on the internet in various formats. At the end of 2013 approximate 2.7 Zeta Byte of data exist in the digital world and approximate 30 petabyte of data is in on process daily. According to evaluation the amount of commercial data worldwide, across all industries, doubles every year [2].

2.1 Structured Data

Data having well defined format and length is generally considered as structured data. Strings, dates, numbers come under this category. These type of data stored in database in rows and columns format. It is the data which used to do transactions using query language from the traditional data sources.

In big data structured data is taking new shape due to technology evolution which provides the new sources of structured data being produced in large volume and in real time through unconventional methods. Some portion of data may be generated from machine and some may be from human machine interaction.

2.2 Unstructured Data

Data that exist in untraditional way is unstructured data and there may be various sources of its generation. It can be generated by humans or a range of devices. Unstructured data is the data we encounter more often. It can be sound, image or video etc. Such types of data cannot be stored in conventional databases. The growth of unstructured data is many times faster than structured data. Unstructured data may contain information that could be of great interest for an organization for better decisions. It is possible to develop customized tools using hardware and software solutions for extracting information from unstructured data.

3. Analysis of Big Data Implications

The growth trend of data and daily advancement in new technologies like data mining, Cloud Computing etc. has started to worry computer professional and researchers that what is the future of big data. As we have already discussed that globally Zeta bytes of data is already exiting and approximate peta bytes of data is uploaded regularly. We have analyzed the impact of various norms which is the real issue of worry to monitor the huge amount of data.

3.1 Impact of Social Network

It always seems good that when somebody provides us free sites and space to upload our videos, pictures, data and share interests, activities, backgrounds, or real-life connections, but it has started giving worry to some group of professionals who are maintaining your free data. They are thinking that how much time must be allowed to all the users to upload what they want to do. As per the third quarter 2013 summary report of facebook [3],the total number of monthly active users on facebook are 700 million plus and it will cross 1000 million within a year or two. With the mobile technology advancements the mobile users are also increasing day by day. We have done our analysis on five most popular social network sites and the analysis of their monthly visited users is as shown in figure 1.

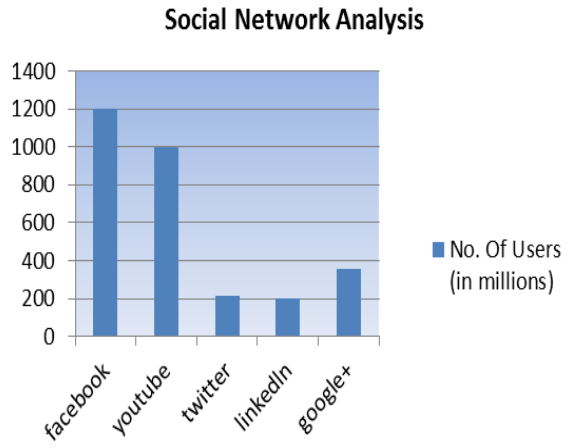

Figure 1: Monthly Active Users of popular Social Sites

3.2 Impact of Search Engines

Search Engines are always used to extract some valuable information, but whether the information is genuine or not that is a big question. There is huge information links

are available on internet which are totally outdated from market and hardly exits. For example we have tried to access "floppy disk" information on one of popular search engine Google and we have found 2,670,000 results in 0.21 seconds [4]. In another example the "Basic language" has approximate 320,000 results. The Big question is that why we are maintaining the huge information which is outdated in market. The increase of traffic is also a big issue to fix the servers limit. The approximate number of users accessing the five most popular search engine monthly in 2013 [Alexa Ranking] is as shown in figure 2.

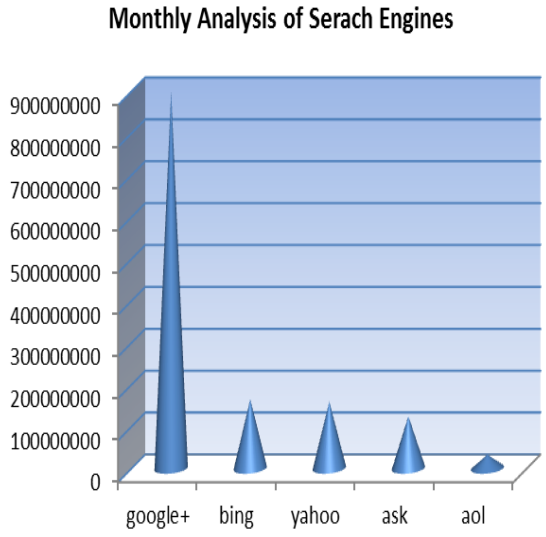

Figure 2: Analysis of Search Engines

3.3 Impact of Private Data

Many organizations have their own data warehouse for their own data or the clients attached with them. This data is not available for public access of Internet users but it is very much part of the global Big Data. Almost all the organizations have their own setup to secure this data and the day to day activities to maintain transactional data. The companies are using both the centralized and distributed approach to maintain their data. The three main factors behind the increase in data for all organizations are:

- Scalability
- Flexibility
- Expenditure

Today almost all the large organizations have more than Exabyte's of private data that is for their own customers. So the challenge starts from this daily growing rate of private data. As the Big data deals with four Vs volume, velocity, variety and veracity maintaining these V's for organizations will always remain challengeable.

4. Opportunities, Challenges and Possible Solutions

It is difficult to identify "totally new" issues brought about by big data. Nonetheless, there are always important aspects to which one hopes to see greater attention and efforts channeled. First, although we have always been trying to handle (increasingly) big data, we have usually assumed that the core computation can be held in memory seamlessly. Sometimes the current data size reaches to such a scale that the data becomes hard to store and even hard for multiple scans. However, many important learning objectives or performance measures are non-linear, non-smooth, non-convex and non-decomposable over samples. For example, AUC (Area Under the ROC Curve) [5], and their optimizations, inherently require repeated scans of the entire dataset. Is it learnable by scanning the data only once, and if it needs to store something, the storage requirement is small and independent to data size? We call this "one-pass learning" and it is important because in many big data applications, the data is not only big but also accumulated over time, hence it is impossible to know the eventual size of the dataset. Fortunately, there are some recent efforts towards this direction, including [6]. On the other hand, although we have big data, are all the data crucial? The answer is very likely that they are not. Then, the question becomes can we identify valuable data subsets from the original big dataset?

Secondly, a benefit of big data to machine learning lies in the fact that with more and more samples available for learning, the risk of over fitting becomes smaller. We all understand that controlling over fitting is one of the central concerns in the design of machine learning algorithms as well as in the application of machine learning techniques in practice. The concern with over fitting led to a natural favor for simple models with less parameters to tune. However, the parameter tuning constraints may change with big data. We can now try to train a model with billions of parameters, because we have sufficiently big data, facilitated by powerful computational facilities that enable the training of such models. The great success of deep learning [7] during the past few years serves as a good showcase. However, most deep learning work strongly relies on engineering tricks that are difficult to be repeated and studied by others, apart from the authors themselves. It is important to study the mysteries behind deep learning; for example, why and

when some ingredients of current deep learning techniques, e.g., pre-training and dropout, are helpful and how they can be more helpful? There have been some recent efforts in this direction [8], [9], [10].

Moreover, we might ask if it is possible to develop a parameter tuning guide to replace the current almost-exhaustive search. Third, we need to note that big data usually contains too many "interests", and from such data we may be able to get "anything we want"; in other words, we can find supporting evidence for any argument we are in favor of. Thus, how do we judge/evaluate the "findings"? One important solution is to turn to statistical hypothesis testing. The use of statistical tests can help at least in two aspects: First, we need to verify that what we have done is really what we wanted to do.

Second, we need to verify that what we have attained is not caused by small perturbations that exist in the data, particularly due to the non-thorough exploitation of the whole data. Although statistical tests have been studied for centuries and have been used in machine learning for decades, the design and deployment of adequate statistical tests is non-trivial, and in fact there have been misuses of statistical tests [11]. Moreover, statistical tests suitable for big data analysis, not only for the computational efficiency but also for the concern of using only part of the data, remain an interesting but under-explored area of research. Another way to check the validity of the analysis results is to derive interpretable models. Although many machine learning models are black-boxes, there have been studies on improving the comprehensibility of models such as rule extraction [12].

Visualization is another important approach, although it is often difficult with dimensions more than three. Moreover, big data usually exists in a distributed manner; that is, different parts of the data may be held by different owners, and no one holds the entire data. It is often the case that some sources are crucial for some analytics goal, whereas some other sources pose less importance. Given the fact that different data owners might warrant the analyzer with different access rights, can we leverage the sources without access to the whole data? What information must we have for this purpose? Even if the owners agree to provide some data, it might be too challenging to transport the data due to its enormous size. Thus, can we exploit the data without transporting them? Moreover, data at different places may have different label quality, and may have significant label noise, perhaps due to crowdsourcing. Can we do learning with low quality and/or even contradictory label information? Furthermore, usually we assume that the data is identically and independently distributed; however, the fundamental assumption can hardly hold across different data sources. Can we learn effectively and efficiently beyond the assumption? There are a few preliminary studies on these important issues for big data, including [13], [14], [15]. In addition, given the same data, different users might have different demands. For example, for product recommendation, some users might demand that highly recommended items are good, and some users might demand that all the recommended items are good; while other users might demand all the good items have been returned. The computational, and storage loads of big data may be inhibitors to the construction of a model for each of the various demands separately. Can we build one model (a "general model" which can be adapted to other demands with cheap minor modifications) to satisfy the various demands? Some efforts have been reported recently in [16].

Another issue related with big data is the increase in temperature of earth. The global mean annual average temperature of earth in 1990 decade was 14.31˚C and approximately 14.51 ˚C in 2000 decade. We are not saying that the big data and the computer technologies are only involved is responsible for increase in temperature, but it plays a vital role for these circumstances. As such to invest huge amount to protect the data warehouses from heat and to avoid heavy electricity load the companies are now shifting to those areas where the temperature remains less than zero degree. But main question is that how long we can do this practice. The possible solutions are that to categorize the valuable and invaluable data. Many social sites are maintaining our many mails, images, and videos etc. that were 10 year old and may be not in use of us. So the essential steps have to be carried out to make some policies to monitor the data storage to protect from global warming.

5. Conclusion

In this paper, we have analyzed the basic concept, characteristics & need of Big Data. The impact of social sites, private data and global warming factors are discussed in this paper. Our analysis illustrate that the growth trend of big data is because of unwanted and unstructured data. The regular monitoring and regular deletion of unwanted and duplicated data are few possible solutions to control the growth rate of big data. Security, storage, searching and environmental changes are the biggest issues related with Big Data and they must be handled carefully in order extract maximum benefit from the big data.

References

[1] Kaisler, S., W. Money, and S. J. Cohen. 2012. "A Decision Framework for Cloud Computing", 45th Hawaii International Conference on System Sciences, Grand Wailea, Maui, HI, Jan 4-7, 2012.

[2] Shiwen Mao · Yunhao Liu "Big Data: A Survey" Min Chen · Published online: 22 January 2014 © Springer Science+Business Media New York 2014.

[3] World's data will grow by 50X in next decade, http://www.computerworld.com/s/article/9217988/World-s-data-will-grow-by-50X-in- next-decade-IDC-study-predicts.

[4] IDC Releases First Worldwide Big Data Technology and Services Market Forecast, Shows Big Data as the Next Essential Capability and a Foundation for the Intelligent Economy http://outsourcing.ulitzer.com/node/2195534.

[5] J. A. Hanley and B. J. McNeil, "A method of comparing the areas under receiver operating characteristic curves derived from the same cases," Radiology, vol. 148, no. 3, pp. 839–843, 1983.

[6] W. Gao, R. Jin, S. Zhu, and Z.-H. Zhou, "One-pass AUC optimization," in Proceedings of the 30th International Conference on Machine Learning, Atlanta, GA, 2013, pp. 906–914.

[7] Y. Bengio, "Learning deep architectures for AI," Foundations and Trends in Machine Learning, vol. 2, no. 1, pp. 1–127, 2009.

[8] P. Baldi and P. J. Sadowski, "Understanding dropout,", in Advances in Neural Information Processing Systems 26. Cambridge, MA: MIT Press, 2013, pp. 2814–2822.

[9] W. Gao and Z.-H. Zhou, "Dropout Rademacher complexity of deep neural networks", CORR abs/1402.3811, 2014.

[10] S. Wager, S. Wang, and P. Liang, "Dropout training as adaptive regularization," Advances in Neural Information Processing Systems 26. Cambridge, MA: MIT Press, 2013, pp. 351–359.

[11] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," Neural Computation, vol. 10, no. 7, pp. 1895 – 1923, 1998.

[12] Z.-H. Zhou, "Rule extraction: Using neural networks or for neural networks?", Journal of Computer Science and Technology, vol. 19, no. 2, 2004, pp. 249–253.

[13] M. Li, W. Wang, and Z.-H. Zhou, "Exploiting remote learners in internet environment with agents," Science China: Information Sciences, vol. 53, no. 1, 2010, pp. 47–76.

[14] M. Mohri and A. Rostamizadeh, "Stability bounds for stationary φ-mixing and β-mixing processes," Journal of Machine Learning Research, vol. 11, pp. 789–814, 2010.

[15] D. Zhou, J. C. Platt, S. Basu, and Y. Mao, "Learning from the wisdom of crowds by minimax entropy," in Advances in Neural Information Processing Systems 25, P. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Cambridge, MA: MIT Press, 2012.

[16] N. Li, I. W. Tsang, and Z.-H. Zhou, "Efficient optimization of performance measures by classifier adaptation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 6, 2013,pp. 1370–1382.

Amarbir Singh is pursuing Ph. D from Punjab Technical University. He has done Master of Computer Applications from Guru Nanak Dev University, Amritsar in 2006 and has published more than ten research papers in various international journals and conferences.

Sarabjit Singh is currently working as assistant professor in Guru nanak Dev University College Verka, Amritsar. He has done Master of Computer Applications from Guru Nanak Dev University, Amritsar in 2005 and has published more than 8 research papers in various international journals and conferences.