International Journal of Advance Research in Science and Engineering
Volume No.06, Special Issue No.(01), Nov 2017
www.ijarse.com
IJARSE
ISSN: 2319-8354

# Investigating the Complexity of Big Data Science

**Dr. Syed Mutahar Aaqib[1], Dr. Kumar Sourabh[2],**

**Syed Ishfaq Manzoor[3]**

[13]*Assistant Professor, Department of Computer Science, Amar Singh College,*
*Cluster University of Srinagar,(India)*
[2] *Assistant Professor, Department of Computer Science, G.G. M. Science College*
*Cluster University of Jammu (India)*

## ABSTRACT

*Nowadays, big data is a flashy and a fancy topic, a trendy research area which everyone seems to be talking about. Independently of what people mean when they use this term, different people view it differently. The challenges this area entails include how to acquire, transfer, store, cleanse, analyze, filter, search, share, and visualize such data. But by just being big is just a matter of volume, velocity, variety and veracity although there is no clear cut agreement in the size threshold where big starts. Indeed, it is easy to acquire large amounts of data, the real goal should not be to just acquire big data and start analyzing and processing it but to ask ourselves this question, for a particular problem, what is the data and how much of it is needed. For some problems this would imply very large sets of data, but for the most of the problems, much less data is what is required essentially. In this paper, we investigate the trade-offs involved in the big data science and the main problems that come with big data viz scalability, redundancy, prejudice, clatter, and privacy issues.*

***Keywords:*** *Big Data, Scalability, Complexity.*

## I. INTRODUCTION

Wikipedia defines big data as a collection of data sets so large and complex that it's difficult to process them using the on-hands existing database management tools or traditional data processing applications. However, what really means on-hand database management tools or traditional data processing applications? Is the volume of data at hand in terabytes or petabytes? In fact, a definition of volume threshold based on current storing and processing capacities might be more reasonable. This definition then may thus be dependent on the machine or the underlying platform. For example, what is big in the mobile world will be smaller in the desktop world. Big data is used in many applications. In the context of the World Wide Web, it can be used to do a simple web search, which will return some information, or to deal with many other data mining problems.

Clearly for a web search, data of a large magnitude is required as we need to look over the whole content over the internet. The crucial difference between a web search and the web data mining is that in the former case we know what we are looking for, while in the latter case the aim is to find something unusual that will be the solution to a (yet) unknown problem. Nowadays, we see a lot of data mining for the sake of it. This has been triggered by the availability of large volumes of data acquired at high velocity [1]. Sometimes it's valid question

to ask ourselves what we actually need in a data set. However, when people over analyze a data set repeatedly just because it is available, newer results become usually less meaningful. In some cases the results also belong to other disciplines (e.g. social sciences) and hence there's no contribution to computer science (CS), but still people tend to publish it in CS circles and academic journals.

Data mining has to be problem driven. And for this we need to answer questions like: What data we need? How much data we need? How the data can be acquired? These days data acquisition might be considered economical and hence big data can be just an object of this step. After the data has been acquired we need to think about transferring and storing it. In fact, transferring one petabyte even over a fast Internet connection (say 100 Mbps) needs more than 12 months, which is not acceptable in most of the real life applications.

Nowadays, numerous organizations pile up hundreds of petabytes of data and process terabytes of it on daily basis.In such a situation, we need to find out that whether all this data is unique or not? Is the source of the data reliable? Is the distribution of data valid or prejudiced? What are the privacy issues? Do we need to keep our data anonymous?

Once all these questions have been dealt with, we need to find out whether we have the capability to process this data?

Is our technique or algorithm scalable? The last and final question that we need to answer is whether all this processing resulted in the production of data that is useful to us and the problem at hand.

Another delicate issue is that most of the time when we need to use big data, the problem is to identify the relevant data pertaining to us inside the large data sets. Many a times this is hard to determine, as we need to discard huge amounts of data, where we have to deal again with prejudice, clutter or spam. Hence, another relevant question is: How we process and filter our data to obtain the relevant data? Hence, working with large volumes of data throws up various challenges related to the issues above. The first one is scalability. Privacy is also very important as it deals with the legal and ethical issues. Other challenges come with the data content and its intrinsic quality, such as redundancy, prejudice, sparsity and clutter. In this article we briefly discuss all these issues. Other aspects of big data like, the heterogeneity of data, are not part of this as it is outside the scope of this article.

## II. SCALABILITY FOR BIG DATA

Scalability is defined as a measure of a system's ability to-without modification-cost-effectively provide increased throughput, reduced response time and/or support more users when hardware resources are added [3]. It also refers to the ability to maintain server's availability, reliability, and performance as the amount of concurrent web requests increase. If an underlying architecture is not able to utilize additional resources to increase the performance, the system is not considered to be scalable [4]. While collecting the data, it is always on our mind that acquiring more if it will fetch us accurate results. But in most of the cases that is far from truth and only results in accumulation of irrelevant data. Thus, More data also implies more noise and clutter.

As the cost of storage and communication bandwidth is getting economically more feasible, scaling-up the communication and hardware of a computer does not imply a proportional increase in cost. But scaling-up a system with commodity hardware is never a lasting solution. Also, the algorithms and the software used to process the data may not scale well. If the algorithm is linear, doubling-up the data, without modifying the core architecture, implies doubling the time. This might still be feasible, but for super-linear algorithms it will not. In such a situation, typical solutions is to scale-out the system by deploying another system in parallel.

As all massive big data solutions run on distributed systems, increasing the quantity of data needs increasing the quantity of host systems, that isn't economical and proportional to the rise required. Another way to deal with this is through the development of faster (appropriate) algorithms, which may not provide accurate insights but will run fast and thereby decreasing the quality of the solution. That is, the time performance improvements should be larger than the loss in the solution quality. This opens a new interesting trade-off challenge in algorithm design and analysis for data mining problems. Another important facet of scalability is the analysis paradigm that is used to speed-up our algorithms. This is dependent, as the degree of parallelization depends on the problem being solved. For instance, not all problems are suitable for the popular map-reduce paradigm [2]. Hence, more research is needed to work out more powerful paradigms, in particular for the analysis of large graphs. In some cases we need to consider the dynamic aspect of big data, as in this case we may need to do online data processing that makes scalability even more difficult. Map-reduce is also not suitable for this case and one on-going initiative for scalable stream data processing is SAMOA [4].

## III. PREJUDICE AND REDUNDANCY IN BIG DATA

Data can be redundant, and most of the times it really is. For instance, in a traffic sensor network that tracks vehicles, there will be a lot of redundant data for all vehicles that are nearby. In the case of the WWW, the lexical redundancy is estimated to be around 25% [9] and semantic redundancy (same meaning but with different words ) makes up an even a larger percentage of the overall web content.

In many situations, while choosing a data sample, the sample may have a specific prejudice. And this prejudice is very difficult to be identified or to be corrected. For example, ‘click data’ in search engines chooses data by ranking and user interface [6, 8]. Hence, Web content is chosen by the ranking function of a particular search engine, which may have a significant impact of the overall quality of the search engines.

Another instance of algorithm prejudice is in hashtag recommendation on social media sites. Imagine that we can recommend hashtags to new objects contributed by people (e.g. social media posts or images). If we do so, in the long run, the recommendation algorithm will generate most hashtags, not the people. Hence, the resulting hash tag space is not entirely the reflection of what the people posted but it also includes data from the machine-algorithm.

## IV. SPARSITY AND CLUTTER IN BIG DATA

A lot of measures within the WWW and alternative forms of knowledge is based on an influence law; therefore mining massive knowledge works all right for the pinnacle of the facility law

distribution with no need a lot of knowledge. This is no longer true once the long tail is taken into account, as a result of the information is sparser. Yet, it typically happens that not enough knowledge covering the long tail is offered once aggregate at the user level. Also, perpetually be cases wherever the most a part of the data distribution will bury the tail (for example, a secondary that means of a question in net search). we tend to explored the scantiness trade-offs relating to personalization and privacy in [5]. Many a times, we invariably attempt to improve results by acquiring more data. While doing so, may not be useful.for instance, if the additional information will increase the clutter in the data, it defeats the basic aim of our solution, as results tend to be inaccurate. In this case, we tend to conjointly reach a saturation level while not seeing any enhancements, thus during this case a lot of data is trashy. Inaccurate results can even result due to internet spam. That is, data fabricated or produced by users in the shape of content , hashtags or links that is aimed at influencing other users within the internet. A common instance of internet spam to boost the ranking of a given web site in a search engine and there are numerous techniques to affect it [7]. And this manipulation can be done at all levels, from increasing a product ratings to even Google Scholar citation counts [7]. Thus, Filtering spam is a complex problem and can be a source of prejudiced data for the acquisition of data.

## V. PERSONALIZATION AND PRIVACY IN BIG DATA

Nowadays, most of the establishments that use personal information for users are bound not to share it with third parties. They additionally use the maximum amount secure communication and storage so that the private data of users is secure and is not purloined. In case of search engines, they have devised information retention policies to assure all the stake holders of the private data, that they fulfill all the legal privacy laws. For instance, they trash the usage logs of even anonymous users every once six months.Like in other cases too, this privacy concern keeps rising, particularly with the arrival of social networks. Generally anonymizing private information isn't enough. For example, in case of the search engines, users are particular about not sharing their personal information like interests, temperament, tastes etc. This includes sexual preferences, health problems or perhaps some on the face of it minor details like hobbies or style in movies that they could not be comfysharing with everyone.

## VI. CONCLUSION

Today big data is certainly a trendy keyword. For this reason we have explored various fundamental questions that we need to address when handling gaint volumes of data. For the same reason, this area is one of the most popular research area in CS and numerous international conferences are being held across the world to address it. As [10] states, could be a matter of size, efficiency, community, or supply. Only time can answer that.

## REFERENCES

[1.] Z. Edosio. Big Data Paradigm- Analysis, Applications and Challenges. In 13th Engineering and Telecommuncation Conference, Vol: 3 University of Bradford, 2015.

[2.] J. Dean and S, Ghemawat. MapReduce: Simplifed Data Processing on Large Clusters. In Proceedings of 6th Symposium on Operating Systems Design and Implementation (OSDI'04). pp. 137-149, 2004.

[3.] S. M. Aaqib, & Sharma L. (2014). Using a Cluster for Efficient Scalability Evaluation of Multithreaded and Event-Driven Web Servers. Intelligent and Soft Computing. Networking and Informatics, 243: 627-636, Springer, ISSN: 2194-5357.

[4.] Bifet. SAMOA: Scalable Advanced Massive Online Analysis. http:// samoa-project.net/, 2013.

[5.] N. Spirin and J. Han. Survey on web spam detection: principles and algorithms. ACM SIGKDD Explorations Newsletter archive. Volume 13 Issue 2, pp. Dec 2011.

[6.] O. Chapelle and Y. Zhang. A dynamic bayesian network click model for web search ranking In Proceedings of the 18th international conference on world wide web (WWW'09). pp. 1-10, 2009.

[7.] E. Delgado Lopez-Cozar, N. Robinson-Garca, and D. Torres-Salinas. Manipulating Google Scholar Citations and Google Scholar Metrics: simple, easy and tempting. 2012.

[8.] R. Jones, R. Kumar, B. Pang, and A. Tomkins.i know what you did last summer": query logs and user privacy. In CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, pages 909{ 914, New York, NY, USA, 2007. ACM.

[9.] F. Radlinski, P.N. Bennett, and E. Yilmaz. Detecting duplicate web documents using click-through data. In Proceedings of the fourth ACM international conference on Web search and data mining pp. 147{156, 2011.

[10.] J. Surowiecki. The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations. Random House, 2004.