Big Data Analytics: Understanding Threats and Opportunities

Authors
Amit Kumar¹, Dr. Ritu sindhu²

¹,² SGT College of Engineering & Technology, India
Email-amit.kec.2008@gmail.com, ritu.sindhu2628@gmail.com

Abstract
Big data is becoming central point of today’s IT environment many organisation using this huge data collection to analyse and predict various patterns and results they further using these analysis for the betterment of the organisation and society as a whole. But when we start studying big data there are a lot of threats and opportunities attached with that. When we deal with big data we to also deal with very much complex data, a huge scale like tsunami of data and privacy issues are coming in big way when we are dealing big data.

Currently there is a lot of ad-hoc approach from the very beginning of capturing the data till final knowledge representation. There is no unified theory across industry to better understand the data and take out relevant information from mountain of data. Another kind of challenges is to deal with unstructured data like text, blogs, tweets and video and images. But among so many bottlenecks Big data also gives the opportunity to understand and predict from huge amount of data with more accuracy.

This paper discusses various threats and opportunities to big data.

Keywords-- Big Data, data mining, predictive modelling, Data management, Data clustering, text analytics.

INTRODUCTION
The amount of data available has exploded in recent years due to new social behaviour, transformation of society as well as the broad dissemination of software systems. Big Data has become a very important factor for the innovation and growth that is based on break-through technologies such as Cloud Computing, Internet objects and Analytics. Big Data is therefore very important to foster the growth of the productivity. Market of Big data analytics is increasing rapidly figure -1 shows the revenue (FY2013) of 3 major organisations working in the field of big data directly or indirectly. Google alone have earned $ 50+ billion in year 2013 which indicates potential of market. Other organisations like SAS, SAP, SPSS working in analytics domains also growing rapidly.
Fig 1: Revenue of Big data related organisations

Fig 2: Big data analysis process

The analysis of the data of large size involves several distinct phases as shown on the figure-2, each of which presents challenges. A lot of people focus unfortunately only on the analysis phase / modelling. While all phases are crucial, it is of little use without the other phases of the conduct of the analysis of the data. Even in the analysis phase, which has received a lot of attention, there is poorly understood complexity in the framework of groups multi-tenanted or the programs of several users running at the same time. Important challenges extend beyond the analysis phase. Although data mining is a process L-4 (figure 3), but to make optimized and efficient, we must see as complete process of capture and storage of the data in the database upto the generation of final pattern or the creation of useful results.

Challenges to Big Data Analysis:
After you have described the multiple phases in the pipeline ‘analysis Big Data’, let us now turn to the common challenges that underlie much, and sometimes all, of these phases. These are represented as five boxes of the second row of the Fig 2

1. Heterogeneity
When humans consume the information, a great heterogeneity is comfortably tolerated. In fact, feature and richness of natural language can make the valuable depth. However, the algorithms of analysis of the machine expect homogeneous data, and cannot understand the nuance. Accordingly, the data must be carefully structured as a first step in the (or before) the analysis of the data. The structure is likely to be required by many traditional systems of data analysis. However, the design less structured is likely to be more effective for many reasons .The systems are more effective if they can store multiple
elements which are all identical in size and structure. An effective representation, access and analysis of semi-structured data require further work.

2. Scale
Of course, the first thing that someone thinks of Big Data is its size. The management of large and increasingly rapid data volumes has been a difficult issue for many decades. In the past, this challenge has been mitigated by the processors more and more quickly, following Moore's law, provide us with the resources necessary to cope with the increasing volumes of data. But, there is a fundamental change in progress, now the volume of data is scaled faster than the computing resources, and speeds of CPU are more or less static.
In the past, the data processing systems had to worry about parallelism between the nodes of a cluster; now it is a case of the parallelism within a single node. Unfortunately, the parallel processing of data which have been applied in the past for the treatment of data across the nodes are not directly applicable intra-node parallelism, since the architecture is very different; for example, there are a lot more of hardware resources such as the caches of the processor and the memory channels of processor which are shared between the cores in a single node. The second radical change that is underway is the cloud evolution, which now aggregates multiple disparate workloads with varying performance goals (e.g. interactive services demand that the data processing engine return back an answer within a fixed response time cap) into very large clusters.

3. Timeliness
The setbacks of size are the speed. More the whole to deal with the data, the more it will have to analyse. The design of a system that effectively treats the size is also likely to lead to a system that can handle a given size of defined data faster. However, it is not only this speed which is usually meant when we talk about the speed in the context of Big Data. On the contrary, there is a rate of acquisition challenge there are many situations in which the result of the analysis is required immediately. For example, if a credit card transaction fraud is suspected, it should ideally be marked before the transaction is completed - the prevention potentially the transaction to take place at all. Obviously, a comprehensive analysis of the history of the purchase of a user is not likely to be possible in real time. Rather, we must develop the partial results in advance so that a small amount of incremental calculation with new data can be used to arrive at a quick determination.

4. Data Protection
Confidentiality of data is another major concern in the context of Big Data. For example, data obtained from the location based (GPS) services. These new model require a user to share his / her location with the service provider, which has led to problems of confidentiality obviously. Masking of the identity of the user, without hiding its location would be a challenge.
5. Human Collaboration
In spite of the enormous progress made in the computational analysis, there are still many patterns that humans can easily detect but computer algorithms have a difficulty in finding. Ideally, the analysis of Big Data will not do all of computation rather it will be designed explicitly for having a human in the loop. The new sub-domain of the visual analytics attempts to do this, particularly with respect to modelling and analysis phase in the pipeline. There is a similar value to the human intervention at all stages for the conduction of the analysis.

A new popular method to exploit human ingenuity to solve problems through people enriched data is Wikipedia, the online encyclopedia, is perhaps the best known example of the people sourced data. It relies on the information provided by strangers. Most often, what they say is correct. However, we should expect that there may be some people who have other motivations and capabilities some may have a reason to provide false information in a deliberate attempt to mislead. While most of these errors will be detected and corrected by others people from same group only. We need technologies to facilitate this. We also need a framework to use in the analysis of these crowd data with the contradictory statements.

6. Develop a unifying theory of Data Mining:
Currently state of the search technique for data mining is too "ad hoc". Many of the techniques are designed for individual problems, such as classification or grouping or clustering, but there is no unifying theory. However, a theoretical framework that unifies the different tasks of data mining, including clustering, classification, association rules, etc, as well as different approaches data mining (such as statistics, data base systems, etc), would help the field and to provide a basis for future research. There is also a possibility and need for the data mining researchers to resolve some long-standing problems in the statistical research, such as problem of avoiding false correlations. This is sometimes linked to the problem of mining to the "deep knowledge", which is the hidden cause of many observations.

7. Mines complex knowledge of complex data
One important type of complex knowledge is in the form of graphs. Recent research has addressed the theme of the discovery of the graphs and structures pattern from large quantities of data, but clearly, more need to be done. Another form of complexity is from data that are not independent and identically distributed. This problem can occur when the data mining from many relations. Normally the objects of interest are not dependent of each other, and are not of a single type. We need systems of mining data which can soundly mine the rich structure of relationships between objects, such as connected Web pages, social networks, etc yet, another important problem is how to mine non-relational database. A large majority of the data for most organizations is in the form of text, not as the database, and in more complex data formats, including the image, multimedia, and data on the Web. Thus, it is necessary to study the
methods of mining of data which go beyond the classification and clustering. A few interesting questions include how to perform better automatic summarization of the text and the way to recognize the movement of objects and persons from Web and wireless of data logs in order to discover useful spatial and temporal knowledge. There is now a strong need to integrating the data mining and knowledge inference. This is a subject of future importance. In particular, an important area is to integrate the background knowledge in the mining of data. The largest gap between what the data mining systems can do today and what we would like to see they are doing, is that they are unable to relate the results of the mining to the real world that they affect all they can do is to hand over the results back to the user. Doing these interference, and automating the entire loop of data mining, requires the representing and using world knowledge of within the system.

8. Distributed Data Mining and Mining Multi-Agent data
The problem of distributed data mining is very important in reference to network. In a distributed environment (for example a sensor or an IP network), one has distributed probes placed at strategic locations inside the network. The problem here is to be able to correlate the observed data in the different probes, and discover patterns in global data observed at all the different probes. There could be different models of distributed data mining here, but it could involve a NOC(Node of Collection) that collects data on the distributed sites, and another in which all sites are treated equally. The objective here would obviously be to reduce to a minimum the amount of data sent between the different sites essentially, to reduce communication overhead. In the distributed mining, a problem is how to mine across multiple heterogeneous data sources: multi-database and the multi-relational mining.

CONCLUSION
We have entered an era of Big Data. Through a better analysis of large volumes of data that become available, it is possible to make more rapid progress in many scientific disciplines and the improvement of the profitability and success of many organisations. However, many technical challenges described in this paper must be addressed before this potential can be fully realized. The challenges include not only the obvious questions of scale, but also the heterogeneity, the lack of unifying theory of data mining, the mining of data in a network environment, security, the integrity of data and privacy. Distributed data mining, timeliness, at all stages of the pipeline of analysis of data acquisition till the interpretation of the results. These technical difficulties are common to a wide variety of application areas. We have to do fundamental research to address these challenges, if we want to achieve the complete benefits of Big Data.

REFERENCES:
2) McKinsey Global Institute, Big Data: The next frontier for innovation, competition and productivity (June 2011)

3) Article: “Big Data at your Service” (July 2012)

4) McKinsey Global Institute, Big Data: The next frontier for innovation, competition and productivity (June 2011),


6) Drowning in numbers -- Digital data will flood the planet—and help us understand it better. The Economist, Nov 18, 2011.