IJESMR
International Journal OF Engineering Sciences & Management Research

BIG DATA:(CHALLENGES AND OPPORTUNITY)

Kiran Patidar*, Pooja Patidar, Shalini Vyas, Swapnil Jain
CSE Students, Mandsaur Institute of Technology, Revas Devra Road, Mandsaur

ABSTRACT

The amount of available data has exploded in the past years because of new social behaviors, societal transformations as well as the vast spread of software systems. Big Data has become a very important driver for innovation and growth that relies on disruptive technologies such as Cloud Computing, Internet of Things and Analytics. Big Data is thus very important to foster productivity growth in Europe since it is affecting not only software-intensive industries but also public services, for example the health, administration and education sectors. A challenge for Europe is to ensure that software and services providers are able to deliver high quality services along the lines of the fast growing number of services and users. Big Data software and services generate value by supporting an innovative eco-system and by enabling completely new solutions that have not been possible before. The value lies in the applications based on advanced data-analysis on top of more general Big Data layers, semantic abstractions or network and physical objects virtualization.

INTRODUCTION

Big Data has become a major topic in the field of ICT. It is evident that Big Data means business opportunities, but also major research challenges. According to McKinsey & Co1 Big Data is "the next frontier for innovation competition and productivity". The impact of Big Data gives not only a huge potential for competition and growth for individual companies, but the right use of Big Data also can increase productivity, innovation, and competitiveness for entire sectors and economies.

The new aspect of Big Data lies within the economic cost of storing and processing such datasets; the unit cost of storage has decreased by many orders of magnitude, amplified by the Cloud business model, significantly lowering the upfront IT investment costs for all businesses. As a consequence, the "Big Data concerns" have moved from big businesses and state research centers, to a mainstream status. As a consequence, the "Big Data concerns" have moved from big businesses and state research centers, to a mainstream status.

NESSI (Networked European Software and Services Initiative) will address current problems and materialize opportunities associated to the use of Big Data.

BIG DATA CONCEPT

Definitions:

"Big Data" is a term encompassing the use of techniques to capture, process, analyses and visualize potentially large datasets in a reasonable timeframe not accessible to standard IT technologies. By extension, the platform, tools and software used for this purpose are collectively called "Big Data technologies".

The new aspect of Big Data lies within the economic cost of storing and processing such datasets; the unit cost of storage has decreased by many orders of magnitude, amplified by the Cloud business model, significantly lowering the upfront IT investment costs for all businesses. As a consequence, the "Big Data concerns" have moved from big businesses and state research centers, to a mainstream status. When dealing with large volumes of data, it is necessary to distribute data and workload over many servers. New designs for databases and efficient ways to support massively parallel processing have led to a new generation of products like the so called no SQL databases and the Hadoop map-reduce platform.

ECONOMIC AND SOCIAL IMPACT OF BIG DATA

Big Data analytics has started to impact all types of organizations, as it carries the potential power to extract embedded knowledge from big amounts of data and react according to it in real time. We exemplify some of the benefits by exploring the following different scenarios.

- Big Data in **healthcare** is associated with the exploding volume of patient-specific data. A prime example is medical imaging where even small pathological features measuring just a few millimeters can be detected in magnetic resonance imaging and in CT scans.
- Various branches of experimental **science** generate vast volumes of experimental data. Petabytes (PB) of data per day is not uncommon in these fields (e.g. research in particle physics produces vast amounts of experimental data within short time frames). Fulfilling the demands of science requires a new way of handling data.
- New technologies produce massive streams of data in real time and space that along time can make it possible to extract patterns of how the structure and form of the city changes and the way in which citizens behave. In such “**smart cities**”, data gathered by sensors integrated with transport data, financial transactions, location of users, social network interaction will provide an entirely new dimension to thinking about how cities function.

TECHNICAL AND SCIENTIFIC CHALLENGES

Big Data Analytics:

Because the current technology enables us to efficiently store and query large datasets, the focus is now on techniques that make use of the complete data set, instead of sampling. This has tremendous implications in areas like machine learning, pattern recognition and classification, to name a few.

Therefore, there are a number of requirements for moving beyond standard data mining techniques:

- a solid scientific foundation to be able to select an adequate method or design .
- a new algorithm (and prove its efficiency and scalability, etc.)
- a technology platform and adequate development skills to be able to implement it.
- a genuine ability to understand not only the data structure (and the usability for a given processing method), but also the business value.

One of the obstacles to widespread analytics adoption is a lack of understanding on how to use analytics to improve the business. The objects to be modeled and simulated are complex and massive, and correspondingly the data is vast and distributed. At the same time, the modeling and simulation software solutions are expected to be simple and general, built on the solid foundations provided by a few robust computational paradigms and naturally oriented towards distributed and parallel computing. Hence, new methodologies and tools for data visualization and simulation are required.

VISUAL ANALYTICS: HOW WE LOOK AT DATA

Visual analytics aims to combine the strengths of human and electronic data processing. Visualization, whereby humans and computers cooperate through graphics, is the means through which this is achieved. Seamless and sophisticated synergies are required for analyzing Big Data, including a variety of multidimensional , temporal data, and solving temporal problems.

The key features of visual analytics research include:

- Emphasis on data analysis, problem solving, and/or decision making;
- Leveraging computational processing by applying automated techniques for data processing, knowledge discovery algorithms, etc.;
- Active involvement of a human in the analytical process through interactive visual interfaces;
- Support for the provenance of analytical results;
- Support for the communication of analytical results to relevant recipients

As the majority of Big Data is dynamic and temporally referenced, it is necessary to take into account the specifics of time15. In contrast to common data dimensions, which are usually “flat”, time has an inherent semantic structure. By convention , time has a hierarchical system of granularities organized in different calendar systems.

Another key issue is supporting analysis at multiple scales. There is much to do for visual analytics in order to change the traditional practice in analysis, focusing on a single scale. Appropriate scales of analysis are not always clear in advance and single optimal solutions are unlikely to exist. Interactive visual interfaces have a great potential for facilitating the empirical search for the acceptable scales of analysis and the verification of results by modifying the scale and the means of any aggregation.

BIG DATA BUSINESS ECOSYSTEM

"An economic community supported by a foundation of interacting organizations and individuals."

In this regard, an ecosystem does not yet exist in Europe for Big Data. However, from at least one perspective a Big Data ecosystem (removal of 'business' is deliberate) does exist in many industries in a very simple form. For example, within an aerospace 'ecosystem' there will be a vast amount of data used across complex supply chains about materials, construction methods, testing, simulation, and so forth. Such data is moved from system to system in the various processes according to defined standards and within regulation. It is also analysis to establish patterns, failures, standards, and so forth.

Moving from such specific Big Data ecosystems to a Big Data Business Ecosystem will not be a straightforward evolution. The problem for Big Data is small patterns "precisely because so many data can now be generated and processed so quickly, so cheaply, and on virtually anything, the pressure is to be able to spot where the new patterns with real added value lie in their immense databases and how they can best be exploited for the creation of wealth and the advancement of knowledge. Small patterns matter because they represent the new frontier of competition, from science to business, from governance to social policies."

CHALLENGES IN BIG DATA ANALYSIS

We now turn to some common challenges of big data analysis-

a. Heterogeneity and Incompleteness :

When humans consume information, a great deal of heterogeneity is comfortably tolerated. Consider, for example, a patient who has multiple medical procedures at a hospital. We could create one record per medical procedure or laboratory test, one record for the entire hospital stay, or one record for all lifetime hospital interactions of this patient. With anything other than the first design, the number of medical procedures and lab tests per record would be different for each patient. The three design choices listed have successively less structure and, conversely, successively greater variety. Greater structure is likely to be required by many (traditional) data analysis systems.

b. Scale :

Of course, the first thing anyone thinks of with Big Data is its size. After all, the word "big" is there in the very name. Managing large and rapidly increasing volumes of data has been a challenging issue for many decades. First, over the last five years the processor technology has made a dramatic shift - rather than processors doubling their clock cycle frequency every 18-24 months, now, due to power constraints, clock speeds have largely stalled and processors are being built with increasing numbers of cores. In the past, large data processing systems had to worry about parallelism across nodes in a cluster; now, one has to deal with parallelism within a single node. Unfortunately, parallel data processing techniques that were applied in the past for processing data across nodes don't directly apply for intra-node parallelism, since the architecture looks very different; for example, there are many more hardware resources such as processor caches and processor memory channels that are shared across cores in a single node.

c. Timeliness :

The flip side of size is speed. The larger the data set to be processed, the longer it will take to analyze. The design of a system that effectively deals with size is likely also to result in a system that can process a given size of data set faster. However, it is not just this speed that is usually meant when one speaks of Velocity in the context of Big Data. There are many situations in which the result of the analysis is required immediately. For example, if a fraudulent credit card transaction is suspected, it should ideally be flagged before the transaction is completed – potentially preventing the transaction from taking place at all. Obviously, a full analysis of a user's purchase history is not likely to be feasible in real-time. Rather, we need to develop partial results in advance so that a small amount of incremental computation with new data can be used to arrive at a quick determination.

d. Privacy :

The privacy of data is another huge concern, and one that increases in the context of Big Data. For electronic health records, there are strict laws governing what can and cannot be done. For other data, regulations, particularly in the US, are less forceful. However, there is great public fear regarding the inappropriate use of personal data, particularly through linking of data from multiple sources. Managing privacy is effectively both a technical and a sociological problem, which must be addressed jointly from both perspectives to realize the promise of big data. There are many additional challenging research problems. For example, we do not know yet how to share private data while limiting disclosure and ensuring sufficient data utility in the shared data. The existing paradigm of differential privacy is a very important step in the right direction, but it unfortunately reduces information content too far in order to be useful in most practical cases.

e. Human Collaboration :

In spite of the tremendous advances made in computational analysis, there remain many patterns that humans can easily detect but computer algorithms have a hard time finding. Indeed, CAPTCHAs exploit precisely this fact to tell human web users apart from computer programs. Ideally, analytics for Big Data will not be all computational – rather it will be designed explicitly to have a human in the loop. The new sub-field of visual analytics is attempting to do this, at least with respect to the modeling and analysis phase in the pipeline. There is similar value to human input at all stages of the analysis pipeline.

d. System Architecture :

Companies today already use, and appreciate the value of, business intelligence. Business data is analyzed for many purposes: a company may perform system log analytics and social media analytics for risk assessment, customer retention, brand management, and so on. Typically, such varied tasks have been handled by separate systems, even if each system includes common steps of information extraction, data cleaning, relational-like processing (joins, group-by, aggregation), statistical and predictive modeling, and appropriate exploration and visualization tools

With Big Data, the use of separate systems in this fashion becomes prohibitively expensive given the large size of the data sets. The expense is due not only to the cost of the systems themselves, but also the time to load the data into multiple systems. In consequence, Big Data has made it necessary to run heterogeneous workloads on a single infrastructure that is sufficiently flexible to handle all these workloads. The challenge here is not to build a system that is ideally suited for all processing tasks. Instead, the need is for the underlying system architecture to be flexible enough that the components built on top of it for expressing the various kinds of processing tasks can tune it to efficiently run these different workloads.

CONCLUSION

We have entered an era of Big Data. Through better analysis of the large volumes of data that are becoming available, there is the potential for making faster advances in many scientific disciplines and improving the profitability and success of many enterprises. However, many technical challenges described in this paper must be addressed before this potential can be realized fully. The challenges include not just the obvious issues of scale, but also heterogeneity, lack of structure, error-handling, privacy, timeliness, provenance, and visualization, at all stages of the analysis pipeline from data acquisition to result interpretation. These technical challenges are common across a large variety of application domains, and therefore not cost-effective to address in the context of one domain alone. Furthermore, these challenges will require transformative solutions, and will not be addressed naturally by the next generation of industrial products. We must support and encourage fundamental research towards addressing these technical challenges if we are to achieve the promised benefits of Big Data.

ACKNOWLEDGE

We sincerely feel oblidge by the guidance provided by our beloved teacher Ms. Priyanka Mangal.

REFERENCE

1. Big data: James Manyika, Michael Chui, Brad Brown.
2. The Search for Analysts to Make Sense of Big Data. Yuki Noguchi.
3. http://www.sdss3.org/collaboration/description
4. “IBM what is a big data? – Bringing big data to the enterprise”
5. Laney , Douglas. “The importance of big data: a definition : Gartner retrieved 21 June 2012.
6. The Age of Big Data. Steve Lohr. *New York Times*, Feb 11, 2012.