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

IJCSMC, Vol. 4, Issue. 6, June 2015, pg.27 – 35

RESEARCH ARTICLE

BIG DATA: OPPORTUNITIES AND CHALLENGES

Rishabh Mishra[1], Dr. Rakesh Sharma[2]

[1]M-Tech Student, [2]Associate Professor, Noida International University, UP

ABSTRACT: In recent years, the rapid development of Internet, Internet of Things, and Cloud Computing have led to the explosive growth of data in almost every industry and business area. Big data has rapidly developed into a hot topic that attracts extensive attention from academia, industry, and governments around the world. In this position paper, we first briefly introduce the concept of big data, including its definition, features, and value. We then identify from different perspectives the significance and opportunities that big data brings to us. Next, we present representative big data initiatives all over the world. We describe the grand challenges (namely, data complexity, computational complexity, and system complexity), as well as possible solutions to address these challenges. Finally, we conclude the paper by presenting several suggestions on carrying out big data projects.

Keywords: Big data, Data complexity, Computational complexity, System complexity.

1. Introduction

In recent years, big data has rapidly developed into a hotspot that attracts great attention from academia, industry, and even governments around the world. Nature and Science have published special issues dedicated to discuss the opportunities and challenges brought by big data and Some even say that big data can be regarded the new petroleum that will power the future information economy. In short, the era of big data has already been in the offing.

What is big data? So far, there is no universally accepted definition. In Wikipedia, big data is defined as "an all-encompassing term for any collection of data sets so large and complex that it becomes difficult to process using traditional data processing applications". From a macro perspective, big data can be regarded as a bond that subtly connects and integrates the physical world, the human society, and cyberspace. Here the physical world has a reflection in cyberspace, embodied as big data, through Internet, the Internet of Things, and other information

© 2015, IJCSMC All Rights Reserved

technologies, while human society generates its big data-based mapping in cyberspace by means of mechanisms like human–computer interfaces, brain–machine interfaces, and mobile Internet [9]. In this sense, big data can basically be classified into two categories, namely, data from the physical world, which is usually obtained through sensors, scientific experiments and observations (such as biological data, neural data, astronomical data, and remote sensing data), and data from the human society, which is often acquired from such sources or domains as social networks, Internet, health, finance, economics, and transportation.

Compared to traditional data, the features of big data can be characterized by 5V, namely, huge Volume, high Velocity, high Variety, low Veracity, and high Value. Actually, the real challenges center around the diversified data types (Variety), timely response requirements (Velocity), and uncertainties in the data (Veracity). Because of the diversified data types, an application often needs to deal with not only traditional structured data, but also semi-structured or unstructured data (including text, images, video, and voice). Timely responses are also challenging because there may not be enough resources to collect, store, and process the big data within a reasonable amount of time. Finally, distinguishing between true and false or reliable and unreliable data is especially challenging, even for the best data cleaning methods to eliminate some inherent unpredictability of data.

2. Significance of big data

Due to its great value, big data has been essentially changing and transforming the way we live, work, and think. In what follows, we describe in detail the significance of big data in various perspectives.

2.1. Significance to national development

At present, the world has completely entered the era of the information age. The extensive use of Internet, Internet of Things, Cloud Computing, and other emerging IT technologies has made various data sources increasing at an unprecedented rate, while making the structures and types of data increasingly complex. Depth analysis and utilization of big data will play an important role in promoting sustained economic growth of countries and enhance the competitiveness of companies.

In the future, big data will become a new point of economic growth. With big data, companies will upgrade and transform to the mode of Analysis as a Service (AaaS), thereby changing the ecology of the IT and other industries. In this context, the global giants of the IT industry (such as IBM, Google, Microsoft, and Oracle) have already begun their technical development planning in the big data era.

At the national level, the capacity of accumulating, processing, and utilizing vast amounts of data will become a new landmark of a country's strength. The data sovereignty of a country in cyberspace will be another great power-game space besides land, sea, air, and outer spaces.

The Big Data Research and Development Initiative, announced by the United States in March 2012, is not only a strategic plan that promotes the US to continuously lead in the high-tech

© 2015, IJCSMC All Rights Reserved

fields, but also a plan to protect its national security and advance its socio-economic development.

In general, the Western countries, represented by the United States, are moving under their national agenda towards a modernization of their national strength through big data research and applications. It is anticipated that future economic and political competitions among countries will be based on exploiting the potential of big data, among other traditional aspects. In short, the research and applications of big data are of strategic importance and significance for improving the competitiveness of any country.

2.2. Significance to industrial upgrades

Big data is currently a common problem faced by many industries, and it brings grand challenges to these industries' digitization and informationization. Research on common problems of big data, especially on breakthroughs of core technologies, will enable industries to harness the complexity induced by data interconnection and to master uncertainties caused by redundancy and/or shortage of data. Everyone hopes to mine from big data demand-driven information, knowledge and even intelligence and ultimately taking full advantage of the big value of big data. This means that data is no longer a byproduct of the industrial sector, but has become a key nexus of all aspects. In this sense, the study of common problems and core technologies of big data will be the focus of the new generation of IT and its applications. It will not only be the new engine to sustain the high growth of the information industry, but also the new tool for industries to improve their competitiveness.

For example, in recent years, cloud computing has rapidly evolved from a vague concept in the beginning to a mature hot technology. Many big companies, including Google, Microsoft, Amazon, Facebook, Alibaba, Baidu, Tencent, and other IT giants, are working on cloud computing technologies and cloud-based computing services. Big data and cloud computing is seen as two sides of a coin: big data is a killer application of cloud computing, whereas cloud computing provides the IT infrastructure to big data. The tightly coupled big data and cloud computing nexus are expected to change the ecosystem of Internet, and even affect the pattern of the entire information industry.

2.3. Significance to scientific research

Big data has caused the scientific community to re-examine its methodology of scientific research and has triggered a revolution in scientific thinking and methods.

It is well-known that the earliest scientific research in human history was based on experiments. Later on, theoretical science emerged, which was characterized by the study of various laws and theorems. However, because theoretical analysis is too complex and not feasible for solving practical problems, people began to seek simulation-based methods, which led to computational science.

The emergence of big data has spawned a new research paradigm; that is, with big data, researchers may only need to find or mine from it the required information, knowledge and

© 2015, IJCSMC All Rights Reserved

intelligence. They even do not need to directly access the objects to be studied. In 2007, the late Turing Award winner, Jim Gray, depicted in his last speech the fourth paradigm of data-intensive scientific research, which separates data-intensive science from computational science.

2.4. Significance to emerging interdisciplinary research

Big data technologies and the corresponding fundamental research have become a research focus in academia. An emerging interdisciplinary discipline called data science has been gradually coming into place. This takes big data as its research object and aims at generalizing the extraction of knowledge from data. It spans across many disciplines, including information science, mathematics, social science, network science, system science, psychology, and economics. It employs various techniques and theories from many fields, including signal processing, probability theory, machine learning, statistical learning, computer programming, data engineering, pattern recognition, visualization, uncertainty modeling, data warehousing, and high performance computing.

Lots of universities and research institutes have even set up undergraduate and/or postgraduate courses on data analytics for cultivating talents, including data scientists and data engineers.

2.5. Significance to helping people better perceive the present

Big Data, especially big networked data, contains a wealth of societal information and can thus be viewed as a network mapped to society. To this end, analyzing big data and further summarizing and finding clues and laws it implicitly contains can help us better perceive the present.

Deep mining information contained in big data can also help people make better decisions. For example, in the presidential election of the United States in November 2012, Barack Obama's campaign team helped Obama by analyzing big data in order to beat Romney and to get re-elected. In the eighteen months before Election Day, Obama's data analysis team created a huge data processing system. Through real-time data collection and analysis, not only could it tell the campaign team how to find voters and to get their attention, but it also analyzed the tendency for voters to vote. Every night, the data analysis team conducted simulation on the election and presented simulation results in the next day to help understand the possibility that Obama might win in some areas, based on which the team can allocate resources more precisely. Later facts demonstrated that the data analysis team played a crucial role in Obama's re-election, far beyond people's imagination.

Analyzing and mining big data can also effectively safeguard public security and combat criminal and economic crimes.

2.6. Significance to helping people better predict the future

Through effective integration and accurate analysis on multi-source heterogeneous big data, better predictions of future trends of events can be achieved. It is possible for big data analysis to

© 2015, IJCSMC All Rights Reserved

even promote sustainable developments of society and economy and further give birth to new industries related to data services.

The ability of big network data has been being highly developed and effectively applied in the field of security and military. As an example, as early as in 2010, the United States released a report entitled "Chinese Nuclear Warhead Storage and Handling System", which claimed that the US found nuclear bases of China in areas like Shaanxi, Jiangxi, and Sichuan. The report even presented the names of cities and counties where the nuclear bases were located. This reports caused a sensation at a global scale. Through this report, the 2049 Project Institute of the United States got into public's attention. Founded in Washington, DC, in 2008, this institute makes use of publicly available data and documents (such as journals and conference papers) to analyze and predict security issues in China related to its military and economy. They completed the report through vertical searches, elaborated analysis, and systematic analysis of big data range anti-ship cruise and ballistic-missile strikes.

Big data-based predictive analysis has been applied to address societal issues, including public health and economic development. Ginsberg, *et al*. found that, if the volume of queries submitted to Google and with keywords like "flu symptom" and "flu treatment" increase in a region, then after a few weeks, the number of influenza patients to the emergency rooms of hospitals in the corresponding area will increase accordingly. With this discovery, they will be able to predict outbreaks of influenza and deploy countermeasures in advance.

On economic development, the United Nations recently launched a new project, called Global Pulse, which expects to use big data to promote the development of global economy. The United Nations will conduct the so-called emotional analysis, which makes use of natural-language-processing software to analyze text messages in social networking sites in order to predict societal issues like unemployment rate, spending cuts and disease outbreaks in a given region. Its overall goal is to utilize digital early warning signals to guide assistance projects in advance in order to prevent an area from re-falling into the plight of poverty.

3. International initiatives on big data

Because of the great significance and value of big data, many countries have launched their plans or initiatives on big data related research and applications. In this section, we briefly overview these efforts.

As mentioned in the previous section, in March 2012, the Obama Administration officially launched the Big Data Research and Development Initiative with an investment of more than US$ 200 million. The initiative involves six federal government agencies, namely, the Department of Defense (DoD), Defense Advanced Research Projects Agency (DARPA), Department of Energy (DoE), National Institutes of Health (NIH), National Science Foundation (NSF), and US Geological Survey (USGS).

Besides the United States, Britain, France, Australia, and Japan have also introduced their big data initiatives.

© 2015, IJCSMC All Rights Reserved

In January 2013, the British government announced a big-data plan of £189 million. On one hand, the plan aims to push new opportunities for using big data in commercial enterprises and research institutions. It further supports with capital and policies the development of big data in medical, agricultural, commercial, academic research and other areas.

In February 2013, the French government published the "Digital Roadmap," which invested €11.5 million to support the development of seven future projects, including big data.

In August 2013, the Australian federal government announced the Australian Public Service Big Data strategy. It intends to promote the service reformation of public sectors by making use of big data analysis, developing better public policies and protecting citizen privacy in order to make Australia among the world's most advanced in the big data field.

The Japanese government announced their national big data strategies, "The Integrated ICT Strategy for 2020" and "Declaration to be the World's Most Advanced IT Nation", in 2012 and 2013, respectively. They plan to develop Japan's new national IT strategy with open public data and big data as its core during 2013–2020, and finally promote Japan as a country with the world's highest standards in the extensive use of big data in the information technology industry.

Finally, the European Commission announced Horizon 2020 as their next framework program for research and innovation, which invests about €120 million on big data-related industrial research and applications. The program defines a research and innovation strategy to guide a successful implementation of big data economy, including excellent science, industrial leadership, and societal challenges.

4. Grand challenges of big data

There are many challenges in harnessing the potential of big data today, ranging from the design of processing systems at the lower layer to analysis means at the higher layer, as well as a series of open problems in scientific research. Among these challenges, some are caused by the characteristics of big data, some, by its current analysis models and methods, and some, by the limitations of current data processing systems. In this section, we briefly describe the major issues and challenges.

4.1. Data complexity

The emergence of big data has provided us with unprecedented large-scale samples when dealing with computational problems, although we now have to face far more complex data objects. As aforementioned, the typical characteristics of big data are diversified types and patterns, complicated inter-relationships, and greatly varied data quality. The inherent complexity of big data (including complex types, complex structures, and complex patterns) makes its perception, representation, understanding and computation far more challenging and results in sharp increases in the computational complexity when compared to traditional computing models based on total data. Traditional data analysis and mining tasks, such as retrieval, topic discovery, semantic analysis, and sentiment analysis, become extremely difficult when using big data. At present, we do not have a good understanding on addressing the complexity of big data. For

© 2015, IJCSMC All Rights Reserved

instance, we lack knowledge regarding the laws of distribution and association relationship of big data. We lack deep understanding on the inherent relationship between data complexity and computational complexity of big data, as well as domain-oriented big data processing methods. All these greatly confine our capacity to design highly efficient computational models and methods for solving problems using big data.

A fundamental problem is how to formulate or quantitatively describe the essential characteristics of the complexity of big data. The study on complexity theory of big data will help understand essential characteristics and formation of complex patterns in big data, simplify its representation, get better knowledge abstraction, and guide the design of computing models and algorithms on big data. To do this, we will need to establish the theory and models of data distribution under multi-modal interrelationships. We will also need to sort out intrinsic connections between data complexity and spatio-temporal computational complexity. Moreover, by modeling and analyzing the intrinsic mechanisms of data complexity, we will be able to expound the principles and mechanisms for processing big data into a solid foundation for big data computing.

4.2. Computational complexity

Three of the key features of big data, namely, multi-sources, huge volume, and fast-changing, make it difficult for traditional computing methods (such as machine learning, information retrieval, and data mining) to effectively support the processing, analysis and computation of big data. Such computations cannot simply rely on past statistics, analysis tools, and iterative algorithms used in traditional approaches for handling small amounts of data. New approaches will need to break away from assumptions made in traditional computations based on independent and identical distribution of data and adequate sampling for generating reliable statistics. When solving problems involving big data, we will need to re-examine and investigate its computability, computational complexity, and algorithms.

New approaches for big data computing will need to address big data-oriented, novel and highly efficient computing paradigms, provide innovative methods for processing and analyzing big data, and support value-driven applications in specified domains. New features in big data processing, such as insufficient samples, open and uncertain data relationships, and unbalanced distribution of value density, not only provide great opportunities, but also pose grand challenges, to studying the computability of big data and the development of new computing paradigms.

To address the computational complexity of big data applications, we will need to focus on the whole life cycle of big data applications in order to study data-centric computing paradigms based on the characteristics of big data. We need to break away from traditional computing-centric paradigms and establish data-centric push-style computing paradigms and explore weak CAP network shared-data system model and its algebraic computational theory. We will need to develop algorithms for distributed and streaming computing and form a big data oriented computing framework where communication, storage, and computing are well integrated and optimized. We will have to study non-deterministic algorithmic theory suitable for big data and depart from the independent-and-identically-distributed assumption made in traditional statistical

© 2015, IJCSMC All Rights Reserved

learning. We also need to explore existing reduction-based computing methods where big data is reduced on demand from being large enough to being just enough, and to being valuable enough. Finally, we will need to develop bootstrapping and sampling based local computation and approximation methods and propose novel theoretical basis for big data algorithms that are scalable to handling large amounts of data.

4.3. System complexity

Big data processing systems suitable for handling a diversity of data types and applications are the key to supporting scientific research of big data. For data of huge volume, complex structure, and sparse value, its processing is confronted by high computational complexity, long duty cycle, and real-time requirements. These requirements not only pose new challenges to the design of system architectures, computing frameworks, and processing systems, but also impose stringent constraints on their operational efficiency and energy consumption.

The design of system architectures, computing frameworks, processing modes, and benchmarks for highly energy-efficient big data processing platforms is the key issue to be addressed in system complexity. Solving these problems can lay the principles for designing, implementing, testing, and optimizing big data processing systems. Their solutions will form an important foundation for developing hardware and software system architectures with energy-optimized and efficient distributed storage and processing.

The evaluation and optimization of energy efficiency of big data processing systems is a great research challenge. Not only do we need to untangle the relationship between complexity and computability of big data applications and between efficiency and energy consumption of processing systems, we will also need to comprehensively measure a variety of energy efficiency factors, including system throughput, parallel processing capabilities, job calculation accuracy, and energy consumption per unit. We also have to take actual workload conditions and scattered and repetitive resources into account. We will need to conduct fundamental research on performance evaluation, distributed system architecture, streaming computing framework, and online data processing, while taking into account features of value sparsity and weak access locality and the life cycle of big data applications. We will need to investigate validation tools, including benchmarks and system performance prediction methods. Through an iterative process of design, implementation, and validation, we will be able to develop big data processing systems with a high data acquisition throughput, low energy consumption, and highly efficient computing.

5. Conclusions

Big data has made a strong impact in almost every sector and industry today. In this paper, we have briefly reviewed the opportunities and significance of big data, as well as some grand challenges that big data brings us. We close by a few suggestions on how to make a big data project successful.

It is no secret that in big data research and applications, industry is ahead of academia. For example, according to the figure Alibaba disclosed in March 2014, their data center has stored

© 2015, IJCSMC All Rights Reserved

more than 100 PB of processed data, which amounts to 100 million high-resolution movies. During the just past "Singles' Day" (also known as "Double 11 Day"), Alibaba pulled in CNY 9.3 billion in sales from this shopping event, which corresponded to around 278 million orders. For this annual shopping event, Alibaba developed a real-time data processing platform called Galaxy, which can handle 5 million transactions per second. The total amount of data that Galaxy can process every day is about 2 PB. Industry is more successful in this respect because it has two essential driving forces: they really need to possess big data in real time and they have the requirements on making better use of the data collected.

The successful applications of big data in industry point to the following necessary conditions for a big data project to be successful. Firstly, there must be very clear requirements, regardless of whether they are technical, social, or economic. Secondly, to efficiently work with big data, we will need to explore and find the kernel structure or kernel data to be processed. Finding kernel data and structures, which are small enough and yet can characterize the behavior and properties of the underlying big data, is non-trivial because it is very domain-specific. Thirdly, a top-down management model should be adopted. Although a bottom-up approach may allow us to solve some niche problems, the isolated solutions often cannot be put together into a complete solution. Finally, the goal should be to solve the entire problem by an integrated solution, rather than striving for isolated successes in a few aspects. In short, an integrated engineering approach should be employed in managing a big-data project.

References

1. V. Mayer-Schonberger, K. CukierBig Data: A Revolution That Will Transform How We Live, Work, and ThinkHoughton Mifflin Harcourt (2013)
2. R. Thomson, C. Lebiere, S. Bennati Human, model and machine: a complementary approach to big data:Proceedings of the 2014 Workshop on Human Centered Big Data Research, HCBDR '14 (2014)
3. A. Cuzzocrea : Privacy and security of big data: current challenges and future research perspectives:Proceedings of the First International Workshop on Privacy and Security of Big Data, PSBD '14 (2014)
4. Big data,Nature, 455 (7209) (2008), pp. 1–136
5. J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, A. Hung in Big data: the next frontier for innovation, competition, and productivity,Tech. rep. McKinsey Global Institute (2011) available at: http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation
6. C. O'Neil, R. Schutt: Doing Data Science: Straight Talk from the Frontline,O'Reilly Media, Inc. (2013)
7. Big data:http://en.wikipedia.org/wiki/Big_data (2014)
8. W.B. Arthur,The second economyavailable at: http://www.images-et-reseaux.com/sites/default/files/medias/blog/2011/12/the-2nd-economy.pdf (2011)
9. T. Kalil:,Big data is a big deal available at: http://www.whitehouse.gov/blog/2012/03/29/big-data-big-deal (2012)
10. T. Hey, S. Tansley, K. Tolle (Eds.), The Fourth Paradigm: Data-Intensive Scientific Discovery, Microsoft Corporation (2009)

© 2015, IJCSMC All Rights Reserved