Information Sciences 275 (2014) 314–347

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Data-intensive applications, challenges, techniques and technologies: A survey on Big Data

C.L. Philip Chen *, Chun-Yang Zhang

Department of Computer and Information Science, Faculty of Science and Technology, University of Macau, Macau, China

ARTICLE INFO

Article history:
Received 28 March 2013
Received in revised form 3 January 2014
Accepted 10 January 2014
Available online 21 January 2014

Keywords:
Big Data
Data-intensive computing
e-Science
Parallel and distributed computing
Cloud computing

ABSTRACT

It is already true that Big Data has drawn huge attention from researchers in information sciences, policy and decision makers in governments and enterprises. As the speed of information growth exceeds Moore's Law at the beginning of this new century, excessive data is making great troubles to human beings. However, there are so much potential and highly useful values hidden in the huge volume of data. A new scientific paradigm is born as data-intensive scientific discovery (DISD), also known as Big Data problems. A large number of fields and sectors, ranging from economic and business activities to public administration, from national security to scientific researches in many areas, involve with Big Data problems. On the one hand, Big Data is extremely valuable to produce productivity in businesses and evolutionary breakthroughs in scientific disciplines, which give us a lot of opportunities to make great progresses in many fields. There is no doubt that the future competitions in business productivity and technologies will surely converge into the Big Data explorations. On the other hand, Big Data also arises with many challenges, such as difficulties in data capture, data storage, data analysis and data visualization. This paper is aimed to demonstrate a close-up view about Big Data, including Big Data applications, Big Data opportunities and challenges, as well as the state-of-the-art techniques and technologies we currently adopt to deal with the Big Data problems. We also discuss several underlying methodologies to handle the data deluge, for example, granular computing, cloud computing, bio-inspired computing, and quantum computing.

1. Introduction

Big Data has been one of the current and future research frontiers. In this year, Gartner listed the "Top 10 Strategic Technology Trends For 2013" [158] and "Top 10 Critical Tech Trends For The Next Five Years" [157], and Big Data is listed in the both two. It is right to say that Big Data will revolutionize many fields, including business, the scientific research, public administration, and so on. For the definition of the Big Data, there are various different explanations from 3Vs to 4Vs. Doug Laney used *volume*, *velocity* and *variety*, known as 3Vs [96], to characterize the concept of Big Data. The term volume is the size of the data set, velocity indicates the speed of data in and out, and variety describes the range of data types and sources. Sometimes, people extend another V according to their special requirements. The fourth V can be *value*, *variability*, or *virtual* [207]. More commonly, Big Data is a collection of very huge data sets with a great diversity of types so that it becomes difficult to process by using state-of-the-art data processing approaches or traditional data processing platforms. In 2012, Gartner retrieved and gave a more detailed definition as: "Big Data are high-volume, high-velocity, and/or high-variety

* Corresponding author.
E-mail addresses: Philip.Chen@ieee.org (C.L. Philip Chen), cyzhangfst@gmail.com (C.-Y. Zhang).

0020-0255/$ - see front matter © 2014 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ins.2014.01.015

information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization". More generally, a data set can be called Big Data if it is formidable to perform capture, curation, analysis and visualization on it at the current technologies.

With diversified data provisions, such as sensor networks, telescopes, scientific experiments, and high throughput instruments, the datasets increase at exponential rate [178,110] as demonstrated in Fig. 1 (source from [67]). The off-the-shelf techniques and technologies that we ready used to store and analyse data cannot work efficiently and satisfactorily. The challenges arise from data capture and data curation to data analysis and data visualization. In many instances, science is legging behind the real world in the capabilities of discovering the valuable knowledge from massive volume of data. Based on precious knowledge, we need to develop and create new techniques and technologies to excavate Big Data and benefit our specified purposes.

Big Data has changed the way that we adopt in doing businesses, managements and researches. Data-intensive science especially in data-intensive computing is coming into the world that aims to provide the tools that we need to handle the Big Data problems. Data-intensive science [18] is emerging as the fourth scientific paradigm in terms of the previous three, namely empirical science, theoretical science and computational science. Thousand years ago, scientists describing the natural phenomenon only based on human empirical evidences, so we call the science at that time as empirical science. It is also the beginning of science and classified as the first paradigm. Then, theoretical science emerged hundreds years ago as the second paradigm, such as Newton's Motion Laws and Kepler's Laws. However, in terms of many complex phenomenon and problems, scientists have to turn to scientific simulations, since theoretical analysis is highly complicated and sometimes unavailable and infeasible. Afterwards, the third science paradigm was born as computational branch. Simulations in large of fields generate a huge volume of data from the experimental science, at the same time, more and more large data sets are generated in many pipelines. There is no doubt that the world of science has changed just because of the increasing data-intensive applications. The techniques and technologies for this kind of data-intensive science are totally distinct with the previous three. Therefore, data-intensive science is viewed as a new and fourth science paradigm for scientific discoveries [65].

In Section 2, we will discuss several transparent Big Data applications around three fields. The opportunities and challenges aroused from Big Data problems will be introduced in Section 3. Then, we give a detailed demonstration of state-of-the-art techniques and technologies to handle data-intensive applications in Section 4, where Big Data tools discussed there will give a helpful guide for expertise users. In Section 5, a number of principles for designing effective Big Data systems are listed. One of the most important parts of this paper, which provides several underlying techniques to settle Big Data problems, is ranged in Section 6. In the last section, we draw a conclusion.

2. Big Data problems

As more and more fields involve Big Data problems, ranging from global economy to society administration, and from scientific researches to national security, we have entered the era of Big Data. Recently, a report [114] from McKinsey institute gives transformative potentials of Big Data in five domains: health care of the United States, public sector administration of European Union, retail of the United States, global manufacturing and personal location data. Their research claims that

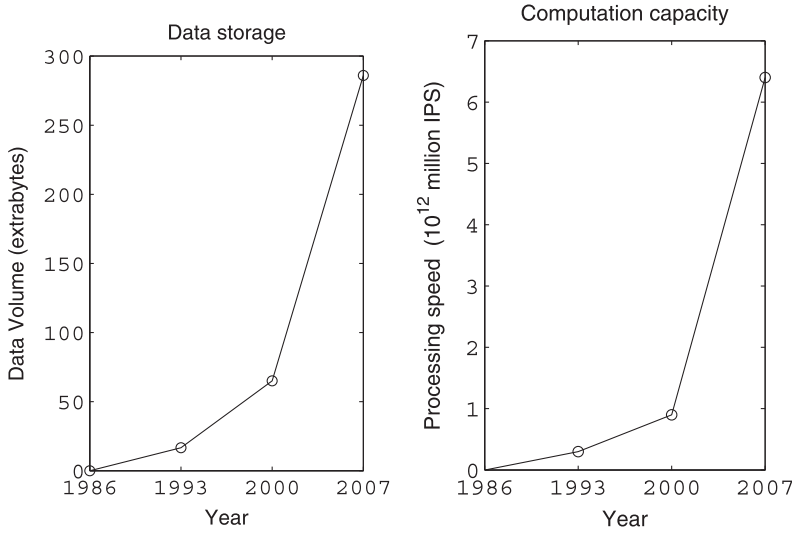

Fig. 1. Data deluge: the increase of data size has surpassed the capabilities of computation.