PROCEEDINGS OF SCIENCE

Challenges of Big Data Analytics

Simon C. Lin[1]
Academia Sinica Grid Computing Centre, (ASGC)
E-mail: Simon.Lin@twgrid.org

Eric Yen
Academia Sinica Grid Computing Centre, (ASGC)
E-mail: Eric.Yen@twgrid.org

The current computer science and industry approach to the Big Data Analytics emphasizes on the importance of Graph processing, how to scale up the capability to process big graph and many algorithms are developed in this line of thinking. However, the more fundamental issue to deal with huge amount of Data objects with many attributes cannot be avoided. Huge amount of datasets from various complex systems are flourishing in the last few years, thus, the exploration of these datasets are supposed to lead the discovery of the unexpected new Data Laws. This paper will examine the challenges of big data, the solution of handling big data and the work has been done in the ASGC.

The International Symposium on Grids and Clouds (ISGC) 2013
March 17-22, 2013
Academia Sinica, Taipei, Taiwan

1 Speaker

PoS(ISGC 2013)025

PROCEEDINGS OF SCIENCE

1. Introduction

The current computer science and industry approach to the Big Data Analytics emphasizes on the importance of Graph processing, how to scale up the capability to process big graph and many algorithms are developed in this line of thinking. However, the more fundamental issue to deal with huge amount of Data objects with many attributes cannot be avoided [1, 2]. This paper will examine the challenges of big data, the solution of handling big data and the work has been done in the ASGC.

2. Challenges of Big Data

The challenges of big data could be summarized in three perspectives, they are: Hardware, Data Deluge and Long-term Preservation.

2.1 Hardware

It is a well-known fact that the economic progress in the past 20 years is driven by the exponential growth in ICT. Figure 1 shows that computer chip capacity doubles every 18 months according to the Moore's Law, data storage doubles every 12 months and communication bandwidth doubles every 9 months. The impact of scaling results in reduced component size and smaller energy consumption. Even so, energy consumption remains a limiting factor of further growth. The energy consumption achieved for today's CPU Floating Point Unit (FPU) is 100 picoJoules (pJ) and in 2018 it has to be reduced to at least ten times less; in addition, energy for DRAM must also be reduced in order to be able to process more data with that system. The electrical transmission is about 2 pJ/bit now; however, the optical transmission could achieve 0.2 pJ/bit. The future CPU must incorporate some kind of photonics. Overall, the computer processing power for Big Data is severely limited by the power consumption even though the scaling to smaller component sizes is possible in principle.

PoS(ISGC 2013)025

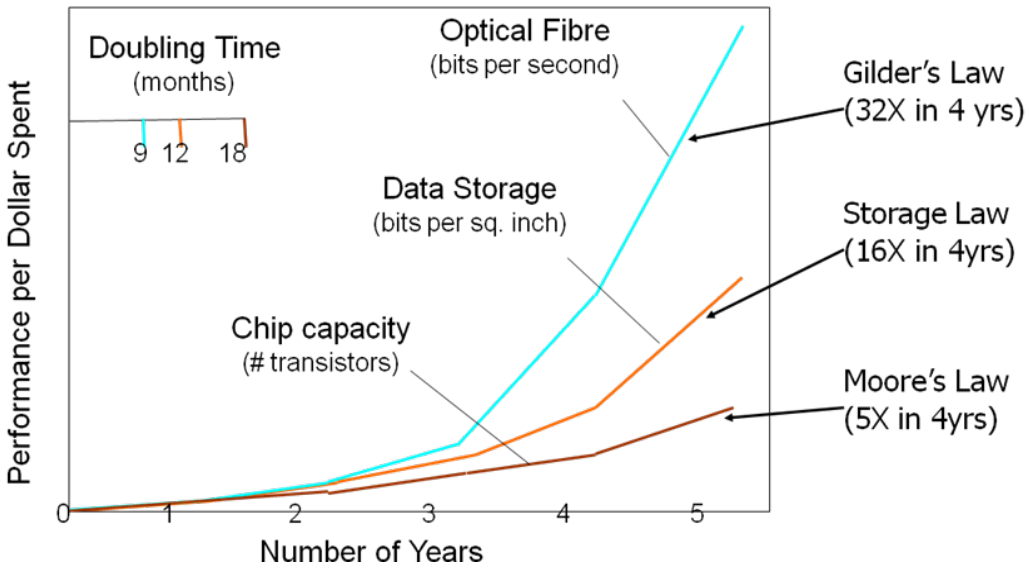

Figure 1: Exponential Growth World

Due to the overheating problem, currently CPU processing speed in GHz (Giga Hertz) has reached a certain limit. Therefore, multi-core with low clock frequency is the mainstream to reduce the power consumption. However, many-cores are not panacea; data movement requires energy, too. If one just looks at the communication part, the electrical connection is 2 picoJoules per bit and that has to be reduced, as mentioned above the optical technology has to be induced apart from some novel architecture, actually, people expect to see one billion ways of parallelism at exascale and this is actually a very daunting task. One might be able to make the hardware components available, but then in order to be able to deliver the performance is another issue. In fact, the commonly agreed 20 MegaWatts power ceiling for the high-end supercomputer imposes limit on the 1 billion processor with the clock cycle around 1 GHz. The HPC seems inevitable to rely on the many cores architecture. One anticipates maybe up to 10 thousand light-weight cores in a CPU node around 2018, but this is only a calculation on the back of the envelope.

The Amdahl's Law is the law for a balanced system. It is not just about the slowest part of the program that determines performance, but also the I/O and the memory laws. In terms of bandwidth, one must be able to process one bit of I/O per instruction and one byte of memory per instruction for the memory law. The typical numbers showed in Figure 2 that modern multi-core systems move away from Amdahl's Law. The ideal situation was that you want to see the number close to 1, but actually they are moving away. The worst is for the planned exascale machine for 2020, the architecture for the exascale machine is actually very imbalanced, they can only process 0.04 bit per flop in terms of I/O and also the memory bandwidth only around 0.01-0.02 Byte per flop which are extremely imbalanced.

PoS(ISGC 2013)025

System	CPU count	GIPS [GHz]	RAM [GB]	diskIO [MB/s]	Amdahl	
					RAM	IO
BeoWulf	100	300	200	3000	0.67	0.08
Desktop	2	6	4	150	0.67	0.2
Cloud VM	1	3	4	30	1.33	0.08
SC1	212992	150000	18600	16900	0.12	0.001
SC2	2090	5000	8260	4700	1.65	0.008
GrayWulf	416	1107	1152	70000	1.04	0.506

Figure 2: Typical Amdahl Numbers (Source: Alexander S. Szalay, Extreme Data-Intensive Scientific Computing)

Apart from expecting exponential growth of the number of cores in the next few years, the existing Programming model does not scale, more innovation required and the cost of re-engineering codes could be considerable. The many-cores CPU architecture may provide opportunity for the flexibility of e-Infrastructure architecture. In order to solve Big Data problem, one must scale out to distributed clouds and scale up to Exascale machines just for its sheer size and the sources and the nature of the geographical distribution of the big Data.

2.2 Data Deluge

The main data deluge problems are twofold: firstly, the underestimation of exponential growth of scientific data and, secondly, the shortage of storage space. According to the media it claims that the whole world generates about 1 Zetta Byte last year (about 150 GB/person) and this has been estimated it will grow at least fifty times by 2020. This number seems not so daunting because definitely personal disk space is usually larger than 150GB each. In fact, this may be a highly conservative estimation because Moore's Law will enable sensors, instruments and detectors to generate unprecedented amount of data in all scientific disciplines. All of these calculations basically have not taken into account of the exponential growth of scientific data.

Figure 3 showed that the total global storage capacity (Hard Disk, NAND, TAPE) shipped in 2011 is 366,400 Peta Byte which is round 0.4EB. The estimated annual increased rate is from 20 to 40%; therefore, in 2020 the total space will be only about 2 Zetta Bytes. However, the data in the year 2020 will reach 50 Zetta Bytes, then we are actually 48 Zetta Bytes in short. So the storage space will grow not as much as we like, this gives a very big constraint on the numbers of data that we can keep on the storage before we can proceed with the processing of Big Data and their Analytics. There will be new algorithmic design issues on what kind of data and on how much data we can keep and then to process and analyze, so this is actually a big problem.

PoS(ISGC 2013)025

	2010	2011
HDD Revenue	$33.5 B	$33.5 B
HDD PB Shipped	330000 PB	330000 PB
HDD $/GB Shipped	$0.10/GB	$0.10/GB
NAND Revenue	$18.5 B	$21.5 B
NAND PB Shipped	10,400 PB	18,600 PB
NAND $/GB	$1.77/GB	$1.16/GB
TAPE LTO Cartridge Revenue	$0.7 B	$0.7B
TAPE LTO Cartridge PB Shipped	15,300 PB	17,800 PB
TAPE LTO Cartridge $/GB	$0.046/GB	$0.038/GB

Figure 3: Global Storage Capacity

Take current data rates for example, the New York Stock Exchange processes about 1.5 TeraBytes of data per day and maintain about 8 PetaBytes; Facebook adds more than 100,000 users, 55M "status" updates and 80M photos daily; Foursquare reports 1.2M location check-ins per week; MEDLINE adds from 1 to 140 publications a day. Those actually will be constrained by how much disk they can buy and then how much space they can own. We only have that much space, although one might be able to treat the data as real-time streaming data and extract some "raw" data for further processing and analysis.

2.3 Long-term Preservation

The final challenge of big data is the long-term preservation. How to keep PetaByte data for a century? Will the format be recognised by then? Are there tools to view, edit and OS available? Any Bit rod? The threats may come from media failure, hardware/software failure, network failure, obsolescence, natural disaster, operator error, external/insider attach, economic failure, organization failure, etc. Bits lost are forever unlike analog materials where contextual information can be used to re-create the original. This is a daunting task where the answer is not generally known yet!

3. How to handle Big Data

Big Data now becomes the most, the hottest keyword being searched for the moment on the Google. The Project of Encyclopedia of DNA Elements (ENCODE) in 2012 has collected 15 TeraBytes and estimated ten-fold of growth every eighteenth month. If coupled with the enormous reduction of the cost of the genome sequencing

PoS(ISGC 2013)025

machines, the cost has dropped from US$3 billion in the year 2000 to now around US$3000 per person. That means lots of this human genome data is going to be recorded. So, in biology the average data growth is ten million times in 14 years. In astronomy, the data volumes of PAN-STARRS reaches 40 PetaBytes; the Square Kilometer Array first light by 2020 with its data volumes of 22,000,000,000 TB per year, which is 700 TB/per second. In climate change, the IPCC in 2012 was 23 PetaBytes and in 2014 the Fifth Assessment Report will reach about 2.5 PetaBytes.

The Big Data in e-Science could be characterized as V^3, Volume, Variety and Velocity (Figure 4). Volume is for sheer size, Variety for different formats of data and for the potentially complexity of the data, Velocity is the speed of data that could be generated.

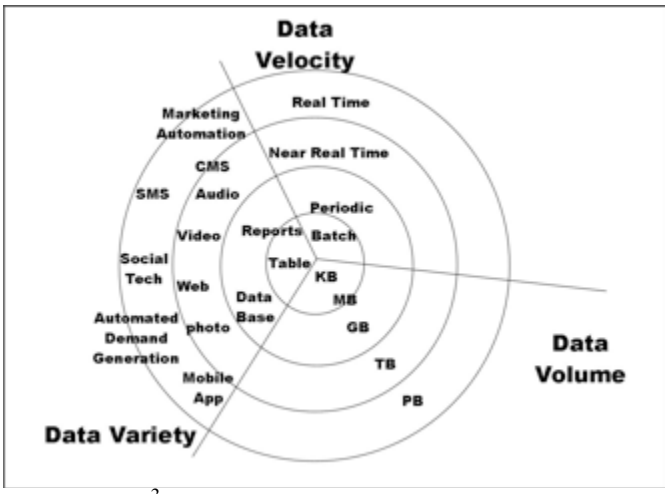

Figure 4: V^3, Volume, Variety and Velocity

Big Data in e-Science reveals a new paradigm, consequently, how to handle big data analytics requires new tools and new thinking. Big Data also changes the nature of scientific computing which now revolving around Data. Science is moving from hypothesis to data driven; in other words, taking the Analysis (Computing) to the Data! Since Big Data in e-Science often involves many disciplines, the analogy of phenomena in different disciplines, data scientists are generally hard to find. It becomes increasingly harder to extract scientific knowledge. Scientists need scale-out solution for analysis, new randomized, incremental algorithms (best result in 1 minute, 1 hour, 1 day, 1 week, etc.), new computational tools and strategies as well as new data intensive scalable architectures. The following section will briefly discussed the Big Data Analytics done at the ASGC.

PoS(ISGC 2013)025

4. Case Study

Huge amount of datasets from various complex systems are flourishing in the last few years, thus, the exploration of these datasets are supposed to lead the discovery of the unexpected new Data Laws. Take medical data as an example, there is a famous Hu Di Ne (Human Disease Network) data from the Harvard Medical School [3]. They took the US Medicare data for the people who are above 65 years old which were about 30 million patients and they try to see different correlations of different diseases with the network analysis. The idea of this is to compare it with the genetic network and see if there is a correlation and also evolutional networkings. The phenotypic disease analysis including Comorbidity study cannot be complete due to the limitation of elderly patient records only.

Collaborating with Taipei Medical University, ASGC has access to the Taiwan National Health Insurance (NHI) records of 23M people of age 0 to 100 from 2000 to 2002. This is a rare and unique dataset for the phenotypic disease analysis due to the non-statistically re-sampled nature and completeness of the population in Taiwan. In fact, this would open up a new opportunity for a truly disease-wise association study (DWAS) [4, 5]. The dataset is big, all pair-wise computation is even bigger. Therefore, the typical Cloud computing technique such as Map-Reduce is employed to generate the necessary dataset for further analysis. Many tools have also been developed to search for new Data Laws, transformation of data, data processing, data analysis and visualization.

The findings are very intriguing. It is found a quantitative method that enables distinguishing of the common and rare diseases which is of great value to the decision support of public health policy. In addition, a new Data Law governs the disease comorbidity for any particular disease is also found. This is very useful in order to compare directly with the data of Genetic Disease Network. Manuscripts are under preparation now and a web site to enable researchers to find all pair-wise diseases comorbidity in the demography of sex and all age groups is also under construction.

Similar works have also been started to study the ancient Chinese text. The ancient Chinese text is troublesome since there is no natural delimiter and no Latin style grammar. However, since language is a function of brain, the study of language as a complex system may reveal the organization of brain eventually. Our approach is based on the idea that semantic structure may be reflected by the text structure, the methods we developed for the Human disease analysis proves to be fruitful in the linguistics case.

The third case is the drug targeting. One often wishes to answer question such as, “Knowing the effectiveness of certain (hundreds to thousands) chemical compounds to a particular protein, what are the other potential compounds from ZINC database of 13M compounds that may also be effective?” Such kind of question actually leads to a

PoS(ISGC 2013)025

new direction that moves away from structure-based to attribute-based drug design which will require making inference from numerous attributes of chemical compounds. We have been making limited progress for the moment, however, as our new theory becomes more complete we believe we will also make substantial progress in this topics.

References

[1] B. Ganter, R. Wille, *Formal Concept Analysis: Mathematical Foundation*, Springer 1999

[2] Z. Pawlak, Rough Sets, *Theoretical Aspects of Reasoning About Data*, Kluwer Academic Publisher 1991

[3] http://hudine.neu.edu/, a site to explore the Human Disease Network.

[4] Goh, K., Cusick, M., Valle, D., Childs, B., Vidal, M., Barabasi, A., *The Human Disease Network*, PNAS, 2007, Vol. 104, No. 21, 8685-8690

[5] Hidalgo, C., Blumm, N., Barabashi, A., Christakis, N., *A Dynamic Network Approach for the Study of Human Phenotypes*, PLoS Computational Biology, 2009, Vol. 5, Issue 4, e1000353

PoS(ISGC 2013)025

