Bilawal Sidhu
WRIT 340
Prof. Warford
July 30, 2013

Demystifying Big Data

As a result of the exponential advancement of technology, our world has experienced an explosion of data. Consumers and businesses alike are creating an inordinate amount of data on a daily basis. To put things in perspective, we live in a world where a whopping 90% of the all data has been generated in the last 4 years [4]. Within these extremely large data sets lay answers to some of our most puzzling commercial, public and scientific questions. The emerging field of big data analysis aims to harness the power of our data in an attempt to obtain actionable insight from it. This article aims to demystify big data by exploring its basic terminology, history and applications.

Introduction

For the past 110 years computing power per dollar has been doubling every year [1]. Even after a glance at the much circulated figure below, exponential growth can be a little difficult to visualize. This is because very few things in life exhibit such a behavior. Picture this— the cellphone in your pocket has more computing power than all of NASA circa 1969, and you got it at a fraction of NASA's $25 billion budget [2]. Intel co-founder Gordon E. Moore famously described this phenomenon wherein he noted that the number of transistors on an integrated circuit could be doubled every 24 months [3]. However, this exponential growth pattern doesn't just apply to CPU's but to most technology such as memory capacity, sensors and the number of pixels in a digital camera [3].

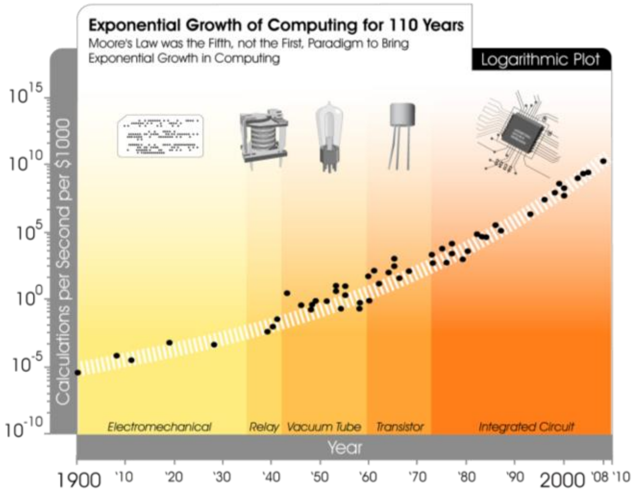

Figure 1: Exponential growth of computing in the last century
Source: http://www.kurzweilai.net/images/exponential_growth_of_computing_110.png

The beauty of the exponential nature of computing is that we presently find ourselves at the far end of the curve, where we are witnessing extremely large leaps in computing power and technological ability. Cheaper computing has resulted in high levels of user adoption among consumers and businesses alike. Computerization is pervasive in the business world, with computers being introduced in more and more areas of the value chain. It is only natural that businesses generate a lot of data. But consumers aren't far behind. Software innovations utilizing faster hardware and global platforms like the internet have allowed for increased *datafication*. This means that things we couldn't quantify or represent in data form before can be done now with the aid of technology. For instance, websites like Facebook and Twitter have allowed us to quantify social interactions and map social networks like never before in the past. We walk around interacting with many internet-connected devices like laptops, cellphones and tablets on a daily basis "communicating, browsing, buying, sharing, searching" and thus creating our own massive trails of data [5]. The exponential nature of technology becomes even more evident when we note that 90% of the world's data has been generated in the last two years [4].

Clearly, big data is around and is here to stay. This article is an attempt to demystify the world of big data as succinctly as possible. We will begin by giving you some background on the history of big data and outline basic terminology. Next, we will explore key driving forces behind the explosion of big data and it's analysis. Finally, we will look at some real-world applications of big data that show case it's immense power.

Big Data and Business Intelligence

We've established that the sheer amount of data in our world is exploding at an exponential rate. Just how big are the datasets that companies like Facebook and Walmart work with? According to the McKinsey Global Institute, the average size of datasets "in many sectors today will range from a few dozen terabytes to multiple petabytes" [5]. A single petabyte is a thousand terabytes! These are the kinds of datasets people are talking about when they refer to *big data*. Simply put, *big data* refers to these massive "datasets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze" [5].

Businesses have long been analyzing their data through a process known as Business Intelligence, or BI. The objective of BI is to provide software and hardware infrastructure that allows companies to sift through copious amounts of raw data, identify what is relevant and obtain actionable business insight from it [6]. In most companies, data from a plethora of sources is aggregated and stored in *data warehouses*, a concept which has been rapidly adopted by businesses since the 1980's [6]. From here business intelligence applications are used to select data sets, ask it questions and get meaningful answers. This process has worked fairly well for companies for the last 20 years. However, with the recent explosion of data, existing business intelligence infrastructure has started to fall flat on its face when attempting to crunch a typical big data set. So how on earth do we analyze such large amounts of data, which will only continue to grow at an exponential rate?

Driving Innovations

Cheaper storage, faster computing and software innovations have been the key driving forces in emergence of big data. Techniques and technologies have been drawn from a variety of fields including statistics, computer science, applied mathematics and economics to make big data analysis possible [5]. First in line is Apache Hadoop, an open-source software framework that enables big data analysis by distributing the computing workload across a large number of traditional servers [7]. This technology was originally developed at Yahoo and was inspired by some of Google's algorithms to deal with large sets of data [5].

In spite of software innovations, big data analysis on traditional data warehouses has proven to be too slow, even when asking the most basic questions. The clear bottleneck in traditional databases is that the data is eventually read from a traditional disk-drive, which has a slow read and write speed, consequently adding a lag time to your analysis [8]. The need to analyze extremely large sets of data, such as real-time roadway traffic conditions, led to the creation of in-memory analytics, which has been both a software and hardware innovation. In-memory databases eliminate this problem by taking an array of servers and storing all of the data sets persistently in their RAM (Random Access Memory), which is an order of magnitude faster than a traditional disk-drive [8]. The data is then backed up to SSD (Solid State Drives) drives, which are still faster than disk-drives. The fast read and write speed of RAM allows companies like Amazon and T-Mobile to obtain answers from their national data sets in under a second, compared to hours on a traditional data warehouse. For instance, an in-memory database enables banks to sift through over 18 billion ATM transactions in a few seconds as opposed to 9 hours [9].

Additionally, cloud computing (http://illumin.usc.edu/printer/131/cloud-computing/) has also made big data more within reach by allowing smaller businesses and research institutions to "rent" computing power needed for such applications. For example, if a company doesn't have the budget to buy its own in-memory database, it can simply rent it from Amazon Web Services and utilize industry standard software from companies such as SAP and Oracle that will interface with their existing business intelligence infrastructure.

What can it do?

Now that we've seen the engineering innovations that have enabled big data analysis, let us examine some of the ways it's being used in the world today.

The European Organization for Nuclear Research, better known as CERN, is on a quest to solve the ultimate mysteries of physical world using the Large Hadron Collider (LHC). CERN quickly realized that they were facing a big data problem. The LHC has 150 million sensors that deliver data at 40 million times per second, resulting in an entire gigabyte of data being sent and stored every second [10]. That's 30 petabytes of data generated annually that needs to be easily accessible by particle physics labs across the globe. The atom smashing events the scientists are in search for are extremely rare. Thus, a majority of the data collected by CERN is irrelevant. However, this large amount of data must be quickly sifted through to identify these extremely rare events. The open-source nature of the tools used by CERN allows particle physicists from around the world access their data sets and collaborate. Such analysis in the past would have been impossible without the advent of big data.

The mapping of the human genome was once an arduous and expensive task taken upon by the Human Genome project. From 1990 to 2003, this multi-billion dollar project worked to successfully map the first human genome [11]. To put things in perspective, a digitized genome has around 100 gigabytes of data [11]. Today, anyone can have their personal genome mapped in a couple of hours at the cost of a few thousand dollars—the difference is simply phenomenal [12]. Big data analysis hasn't just made it cheaper and easier to decode the human genome, but it has also opened up a whole new range of applications. Scientists can utilize genome data sets to pinpoint anomalies and identify genetically linked conditions [12]. This allows scientists to identify these conditions sooner in a patient's life, so that adequate medical attention can be given before it's too late.

Conclusion

Big data has given consumers and businesses a larger and more detailed picture of our world than ever before. Not only can we view a breadth of information to see overarching trends, we can also drill down into the nitty-gritty to identify the most miniscule fluctuations. Given the exponential growth of technology, the amount of data we generate will continue increase and so will our ability to analyze it. The future holds exiting possibilities by giving us answers to questions that we previous considered unanswerable or beyond our reach. Who knows what answers big data will bring forth about everything ranging from the fabric of our reality to our consumption tendencies?

References

[1] Trace Research and Development Center (2007). *Chart depicting exponential growth of computing power* [Online]. Available: http://trace.wisc.edu/tech-overview/indexe9e5.html?attachment_id=256

[2] NASA (2009). *Do-It-Yourself Podcast: Rocket Evolution* [Online]. Available: http://www.nasa.gov/audience/foreducators/diypodcast/rocket-evolution-index-diy.html

[3] Kurzweil Accelerating Intelligence (2001). *The Law of Accelerating Returns* [Online]. Available: http://www.kurzweilai.net/the-law-of-accelerating-returns

[4] Science Daily (2013). *Big Data, for Better or Worse: 90% of World's Data Generated Over Last Two Years* [Online]. Available: http://www.sciencedaily.com/releases/2013/05/130522085217.htm

[5] McKinsey & Company (2011). *Big data: The next frontier for innovation, competition, and productivity* [Online]. Available: http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

[6] Ryan Mucahy (2012). CIO. *Business Intelligence Definition and Solutions* [Online]. Available: http://www.cio.com/article/40296/Business_Intelligence_Definition_and_Solutions

[7] Cloudera (2013). *Hadoop and Big Data* [Online]. Available: http://www.cloudera.com/content/cloudera/en/why-cloudera/hadoop-and-big-data.html

[8] Sreedhar Kajeepeta (2012). Information Week. *The Ins And Outs Of In-Memory Analytics* [Online]. Available: http://www.informationweek.com/software/business-intelligence/the-ins-and-outs-of-in-memory-analytics/240007541

[9] SAP (2012). *A Primer On Big Data And Its Value* [Online]. Available: http://m.sap.com/content/businessinnovation/en_us/big-data-articles/2012/10/a_primer_on_big_data.html

[10] Rachel Barnes (2013). Marketing Magazine. *Big data issues? Try coping with the Large Hadron Collider* [Online]. Available: http://www.marketingmagazine.co.uk/article/1185012/big-data-issues-try-coping-large-hadron-collider

[11] Jacqueline Vanacek (2012). Forbes. *How Cloud and Big Data are Impacting the Human Genome -- Touching 7 Billion Lives* [Online]. Available: http://www.forbes.com/sites/sap/2012/04/16/how-cloud-and-big-data-are-impacting-the-human-genome-touching-7-billion-lives/

[12] Karlin Lillington (2001). The Irish Times. *What Big Data can tell you about your genome – and why it matters* [Online]. Available: http://www.irishtimes.com/business/sectors/technology/what-big-data-can-tell-you-about-your-genome-and-why-it-matters-1.1379643