2nd International Conference on Current Research Trends in Engineering and Technology
© 2018 IJSRSET | Volume 4 | Issue 5 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section: Engineering and Technology

Study on Big Data Challenges and Opportunities

Meet Hire[1], Mayur Gavande[2], Krupit Chovatiya[3], Aditya Tekale[4], Mr. Pritesh Patel[5]

Dr. Sheshang Degadwala[6]

[1-4]U.G. Student Computer Engineering, Sigma Institute of Engineering, Vadodara, Gujarat, India
[5]Assistant Professor of Computer Department, Sigma Institute of Engineering, Vadodara, Gujarat, India
[6]Head of Department Computer Engineering, Sigma Institute of Engineering, Vadodara, Gujarat, India

ABSTRACT

Big data analytics offers promise in many business sectors, and health care is looking at big data to provide answers to many age-related issues, particularly dementia and chronic disease management. The purpose of this review was to summarize the challenges faced by big data analytics and the opportunities that big data opens in health care. The top challenges were issues of data structure, security, data standardization, storage and transfers, and managerial skills such as data governance. The top opportunities revealed were quality improvement, population management and health, early detection of disease, data quality, structure, and accessibility, improved decision making, and cost reduction.

Keywords: Big Data, Analytics, Health Care, Human Genome, Electronic Medical Record

I. INTRODUCTION

We are awash in a flood of data today. In a broad range of application areas, data is being collected at unprecedented scale. Decisions that previously were based on guesswork, or on painstakingly constructed models of reality, can now be made based on the data itself. Such Big Data analysis now drives nearly every aspect of our modern society, including mobile services, retail, manufacturing, financial services, life sciences, and physical sciences.

Scientific research has been revolutionized by Big Data [CCC2011a]. The Sloan Digital Sky Survey [SDSS2008] has today become a central resource for astronomers the world over. The field of Astronomy is being transformed from one where taking pictures of the sky was a large part of an astronomer's job to one where the pictures are all in a database already and the astronomer's task is to find interesting objects and phenomena in the database. In the biological sciences, there is now a well- established tradition of depositing scientific data into a public repository, and also of creating public databases for use by other scientists. In fact, there is an entire discipline of bioinformatics that is largely devoted to the duration and analysis of such data. As technology advances, particularly with the advent of Next Generation Sequencing, the size and number of experimental data sets available is increasing exponentially.

Big Data has the potential to revolutionize not just research, but also education [CCC2011b]. A recent detailed quantitative comparison of different

approaches taken by 35 charter schools in NYC has found that one of the top five policies correlated with measurable academic effectiveness was the use of data to guide instruction [DF2011]. Imagine a world in which we have access to a huge database where we collect every detailed measure of every student's academic performance. This data could be used to design the most effective approaches to education, starting from reading, writing, and math, to advanced, college-level,courses. We are far from having access to such data, but there are powerful trends in this direction. In particular, there is a strong trend for massive Web deployment of educational activities, and this will generate an increasingly large amount of detailed data about students' performance.

It is widely believed that the use of information technology can reduce the cost of healthcare while improving its quality [CCC2011c], by making care more preventive and personalized and basing it on more extensive (home-based) continuous monitoring. McKinsey estimates [McK2011] a savings of 300 billion dollars every year in the US alone.

In a similar vein, there have been persuasive cases made for the value of Big Data for urban planning (through fusion of high-fidelity geographical data), intelligent transportation (through analysis and visualization of live and detailed road network data), environmental modeling (through sensor networks ubiquitously collecting data) [CCC2011d], energy saving (through unveiling patterns of use), smart materials (through the new materials genome initiative [MGI2011]), computational social sciences.

In 2010, enterprises and users stored more than 13 megabytes of new data; this is over 50,000 times the data in the Library of Congress. The potential value of global personal location data is estimated to be $700 billion to end users, and it can result in an up to 50% decrease in product development and assembly costs, according to a recent McKinsey report [McK2011]. McKinsey predicts an equally great effect of Big Data in employment, where 140,000-190,000 workers with "deep analytical" experience will be needed in the US; furthermore, 1.5 million managers will need to become data-literate. Not surprisingly, the recent PCAST report on Networking and IT R&D [PCAST2010] identified Big Data as a "research frontier" that can "accelerate progress across a broad range of priorities." Even popular news media now appreciates the value of Big Data as evidenced by coverage in the Economist [Eco2011], the New York Times [NYT2012], and National Public Radio [NPR2011a, NPR2011b].

While the potential benefits of Big Data are real and significant, and some initial successes have already been achieved (such as the Sloan Digital Sky Survey), there remain many technical challenges that must be addressed to fully realize this potential. The sheer size of the data, of course, is a major challenge, and is the one that is most easily recognized. However, there are others. Industry analysis companies like to point out that there are challenges not just in Volume, but also in Variety and Velocity [Gar2011], and that companies should not focus on just the first of these. By Variety, they usually mean heterogeneity of data types, representation, and semantic interpretation. By Velocity, they mean both the rate at which data arrive and the time in which it must be acted upon. While these three are important, this short list fails to include additional important requirements such as privacy and usability.

The analysis of Big Data involves multiple distinct phases as shown in the figure below, each of which introduces challenges. Many people unfortunately focus just on the analysis/modelling phase: while that phase is crucial, it is of little use without the other phases of the data analysis pipeline. Even in the analysis phase, which has received much attention, there are poorly understood complexities in the context of multi-tenanted clusters where several users' programs run concurrently. Many significant challenges extend beyond the analysis phase. For example, Big Data has to be managed in context, which may be noisy, heterogeneous and not include an upfront model. Doing so raises the need to track provenance and to handle uncertainty and error: topics that are crucial to success, and yet rarely

mentioned in the same breath as Big Data. Similarly, the questions to the data analysis pipeline will typically not all be laid out in advance. We may need to figure out good questions based on the data. Doing this will require smarter systems and also better support for user interaction with the analysis pipeline. In fact, we currently have a major bottleneck in the number of people empowered to ask questions of the data and analyse it [NYT2012]. We can drastically increase this number by supporting

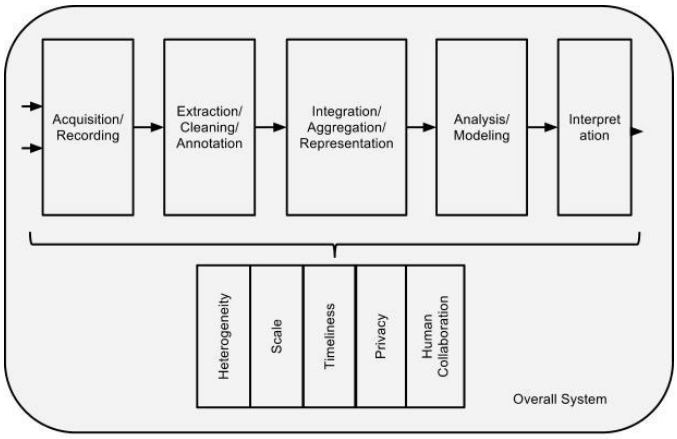

Figure 1: The Big Data Analysis Pipeline.

many levels of engagement with the data, not all requiring deep database expertise. Solutions to problems such as this will not come from incremental improvements to business as usual such as industry may make on its own. Rather, they require us to fundamentally rethink how we manage data analysis.Fortunately,existingcomputational techniques can be applied, either as is or with some extensions, to at least some aspects of the Big Data problem. For example, relational databases rely on the notion of logical data independence: users can think about what they want to compute, while the system (with skilled engineers designing those systems) determines how to compute it efficiently. Similarly, the SQL standard and the relational data model provide a uniform, powerful language to express many query needs and, in principle, allows customers to choose between vendors, increasing competition. The challenge ahead of us is to combine these healthy features of prior systems as we devise novel solutions to the many new challenges of Big Data.

In this paper, we consider each of the boxes in the figure above, and discuss both what has already been done and what challenges remain as we seek to exploit Big Data. We begin by considering

2. Phases in the Processing Pipeline

2.1 Data Acquisition and Recording

Big Data does not arise out of a vacuum: it is recorded from some data generating source. For example, consider our ability to sense and observe the world around us, from the heart rate of an elderly citizen, and presence of toxins in the air we breathe, to the planned square kilometre array telescope, which will produce up to 1 million terabytes of raw data per day. Similarly, scientific experiments and simulations can easily produce pet bytes of data today.

Much of this data is of no interest, and it can be filtered and compressed by orders of magnitude. One challenge is to define these filters in such a way that they do not discard useful information. For example, suppose one sensor reading differs substantially from the rest: it is likely to be due to the sensor being faulty, but how can we be sure that it is not an artifact that deserves attention? In addition, the data collected by these sensors most often are spatially and temporally correlated (e.g., traffic sensors on the same road segment). We need research in the science of data reduction that can intelligently process this raw data to a size that its users can handle while not missing the needle in the haystack. Furthermore, we require “on-line” analysis techniques that can process such streaming data on the fly, since we cannot afford to store first and reduce afterward.

The second big challenge is to automatically generate the right metadata to describe what data is recorded and how it is recorded and measured. For example, in scientific experiments, considerable detail regarding specific experimental conditions and procedures may be required to be able to interpret the results correctly, and it is important that such metadata be recorded

with observational data. Metadata acquisition systems can minimize the human burden in recording metadata. Another important issue here is data provenance. Recording information about the data at its birth is not useful unless this information can be interpreted and carried along through the data analysis pipeline. For example, a processing error at one step can render subsequent analysis useless; with suitable provenance, we can easily identify all subsequent processing that dependent on this step. Thus we need research both into generating suitable metadata and into data systems that carry the provenance of data and its metadata through data analysis pipelines.

2.2 Information Extraction and Cleaning

Frequently, the information collected will not be in a format ready for analysis. For example, consider the collection of electronic health records in a hospital, comprising transcribed dictations from several physicians, structured data from sensors and measurements (possibly with some associated uncertainty), and image data such as x-rays. We cannot leave the data in this form and still effectively analyze it. Rather we require an information extraction process that pulls out the required information from the underlying sources and expresses it in a structured form suitable for analysis. Doing this correctly and completely is a continuing technical challenge. Note that this data also includes images and will in the future include video; such extraction is often highly application dependent (e.g., what you want to pull out of an MRI is very different from what you would pull out of a picture of the stars, or a surveillance photo). In addition, due to the ubiquity of surveillance cameras and popularity of GPS enabled mobile phones, cameras, and other portable devices, rich and high fidelity location and trajectory (i.e., movement in space) data can also be extracted.

We are used to thinking of Big Data as always telling us the truth, but this is actually far from reality. For example, patients may choose to hide risky behaviour and caregivers may sometimes miss-diagnose a condition; patients may also inaccurately recall the name of a drug or even that they ever took it, leading to missing information in (the history portion of) their medical record. Existing work on data cleaning assumes well recognized constraints on valid data or well understood error models; for many emerging Big Data domains these do not exist.

2.3 Data Integration and Aggregation

Given the heterogeneity of the flood of data, it is not enough merely to record it and throw it into a repository. Consider, for example, data from a range of scientific experiments. If we just have a bunch of data sets in a repository, it is unlikely anyone will ever be able to find, let alone reuse, any of this data. With adequate metadata, there is some hope, but even so, challenges will remain due to differences in experimental details and in data record structure.

Data analysis is considerably more challenging than simply locating, identifying, understanding, and citing data. For effective large scale analysis all of this has to happen in a completely automated manner. This requires differences in data structure and semantics to be expressed in forms that are computer understandable, and then "robotically" resolvable. There is a strong body of work in data integration that can provide some of the answers. However, considerable additional work is required to achieve automated error free difference resolution.

Even for simpler analyses that depend on only one data set, there remains an important question of suitable database design. Usually, there will be many alternative ways in which to store the same information. Certain designs will have advantages over others for certain purposes, and possibly drawbacks for other purposes. Witness, for instance, the tremendous variety in the structure of bioinformatics databases with information regarding

substantially similar entities, such as genes. Database design is today an art, and is carefully executed in the enterprise context by highly paid professionals. We must enable other professionals, such as domain scientists, to create effective database designs, either through devising tools to assist them in the design process or through forgoing the design process completely and developing techniques so that databases can be used effectively in the absence of intelligent database design.

2.4 Query Processing, Data Modelling, and Analysis

Methods for querying and mining Big Data are fundamentally different from traditional statistical analysis on small samples. Big Data is often noisy, dynamic, heterogeneous, interrelated and untrustworthy. Nevertheless, even noisy Big Data could be more valuable than tiny samples because general statistics obtained from frequent patterns and correlation analysis usually overpower individual fluctuations and often disclose more reliable hidden patterns and knowledge. Further, interconnected Big Data forms large heterogeneous information networks, with which information redundancy can be explored to compensate for missing data, to crosscheck conflicting cases, to validate trustworthy relationships, to disclose inherent clusters, and to uncover hidden relationships and models.

Mining requires integrated, cleaned, trustworthy, and efficiently accessible data, declarative query and mining interfaces, scalable mining algorithms, and big data computing environments. At the same time, data mining itself can also be used to help improve the quality and trustworthiness of the data, understand its semantics, and provide intelligent querying functions. As noted previously, real life medical records have errors, are heterogeneous, and frequently are distributed across multiple systems. The value of Big Data analysis in health care, to take just one example application domain, can only be realized if it can be applied robustly under these difficult conditions. On the flip side, knowledge developed from data can help in correcting errors and removing ambiguity. For example, a physician may write "DVT" as the diagnosis for a patient. This abbreviation is commonly used for both "deep vein thrombosis" and "diverticulitis," two very different medical conditions. A knowledge base constructed from related data can use associated symptoms or medications to determine which of two the physician meant.

Big Data is also enabling the next generation of interactive data analysis with real time answers. In the future, queries towards Big Data will be automatically generated for content creation on websites, to populate hot lists or recommendations, and to provide an ad hoc analysis of the value of a data set to decide whether to store or to discard it. Scaling complex query processing techniques to terabytes while enabling interactive response times is a major open research problem today.

A problem with current Big Data analysis is the lack of coordination between database systems, which host the data and provide SQL querying, with analytics packages that perform various forms of non SQL processing, such as data mining and statistical analyses. Today's analysts are impeded by a tedious process of exporting data from the database, performing a non SQL process and bringing the data back. This is an obstacle to carrying over the interactive elegance of the first generation of SQL driven OLAP systems into the data mining type of analysis that is in increasing demand. A tight coupling between declarative query languages and the functions of such packages will benefit both expressiveness and performance of the analysis.

II. CONCLUSION

Big data and the use of advanced analytics have the potential to advance the way in which providers leverage technology to make informed clinical decisions. However, the vast amounts of information

generated annually within health care must be organized and compartmentalized to enable universal accessibility and transparency between health care organizations.

Our systematic literature review revealed both challenges and opportunities that big data offers to the health care industry. The literature mentioned the challenges of data structure and security in at least 50% of the articles reviewed. The literature also mentioned the opportunities of increased quality, better management of population health, early detection of disease, and data quality structure and accessibility in at least 50% of the articles reviewed. These findings identify foci for future research.

III. REFERENCES

[1]. PRISMA. [2015-07-30]. Welcome to the Preferred Reporting Items for Systematic Reviews andMeta-Analyses(PRISMA)website! 2009. http://www.prisma-statement.org

[2]. McAfee A, Brynjolfsson E. Big data: the management revolution. Harv Bus Rev. 2012 Oct;90(10):60–6, 68, 128. [PubMed]

[3]. Heudecker N. [2016-11-08]. Hype Cycle for BigData.Gartner.2013Jul31.https://www.gartner.com

[4]. Kayyali B, Knott D, Van Kuiken S. [2016-11-11]. The big-data revolution in US health care: accelerating value and innovation. McKinsey & Company. 2013 Apr. https://digitalstrategy.nl

[5]. Chawla NV, Davis DA. Bringing big data to personalized healthcare: a patient-centered framework. J Gen Intern Med. 2013 Sep;28(Suppl 3):S660–5. doi: 10.1007/s11606-013-2455-8.http://europepmc.org

[6]. US Department of Health and Human Services Genomics Data Sharing. 2014.https://gds.nih.gov

[7]. Feldman B, Martin E, Skotnes T. [2016-11-09]. Big data in healthcare hype and hope. GHDonline. 2012 Oct. https://www.ghdonline.org