BIG DATA: CHALLENGES AND PROSPECTS

Sarika Choudhary[1], Ritika Saroha[2], Yatan Dahiya[3]

[1,2] M.tech (Network Security), School of Engineering & Sciences,
BPS Mahila Vishwavidyalaya, Sonepat, Haryana, (India)
[3]M.Tech (CSE), Baba Mastnath College of Engg., MD University, Rohtak (India)

ABSTRACT

The problems start during data acquisition, when the bulk data requires us to make decisions, currently in an ad hoc manner, about what data to keep and what to discard, and how to store what we keep reliably with the right metadata. Many data today is not natively in structured format, for e.g.: blogs and tweets are weakly structured pieces of text, while images and video are structured for storage and display, but not for semantic content and search, so transforming such content into a structured format for later study is a major test. The objective of this paper is to discuss the characteristics of big data as well as the challenges and opportunities for big data analytics – the process of extracting knowledge from sets of big data.

However, it is hard, requiring us to rethink data analysis systems in fundamental ways. A major speculation in Big Data, properly directed, can result not only in major scientific advances, but also place the foundation for the next generation of advances in science, medicine, and business.

Keywords: Data Acquisition, Big Data Analysis, Heterogeneity, Privacy, Query Processing.

I INTRODUCTION

In a broad range of application areas, data is being collected at extraordinary 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, manufacturing, financial services and life sciences etc.

The field of Astronomy is being changed 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 observable fact in the database.

In the biological sciences, there is now a well-established tradition of depositing technical data into a public repository, and also of creating public databases for use by other scientists. In fact, there is a whole discipline of bioinformatics i.e. largely devoted to the duration and analysis of such data. As technology advances, particularly with the initiation of Next Generation Sequencing, the size and number of experimental data sets available is increasing exponentially.

Imagine a world in which we have access to a huge database where we gather every detailed calculate of every student's academic performance. This data could be used to plan the most helpful approaches to education, starting from reading, writing, and math, to advanced, college-level, courses. We are far from having access to such data. In particular, there is a strong trend for massive Web exploitation of educational activities, and this will produce a gradually larger amount of detailed data about students' performance.

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. 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 various distinct phases as shown in the figure below, each of which introduces challenges. Many people unfortunately focus just on the examination/modeling phase: while that phase is critical, 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 e.g. Big Data has to be managed in context, which may be noisy, heterogeneous. 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.

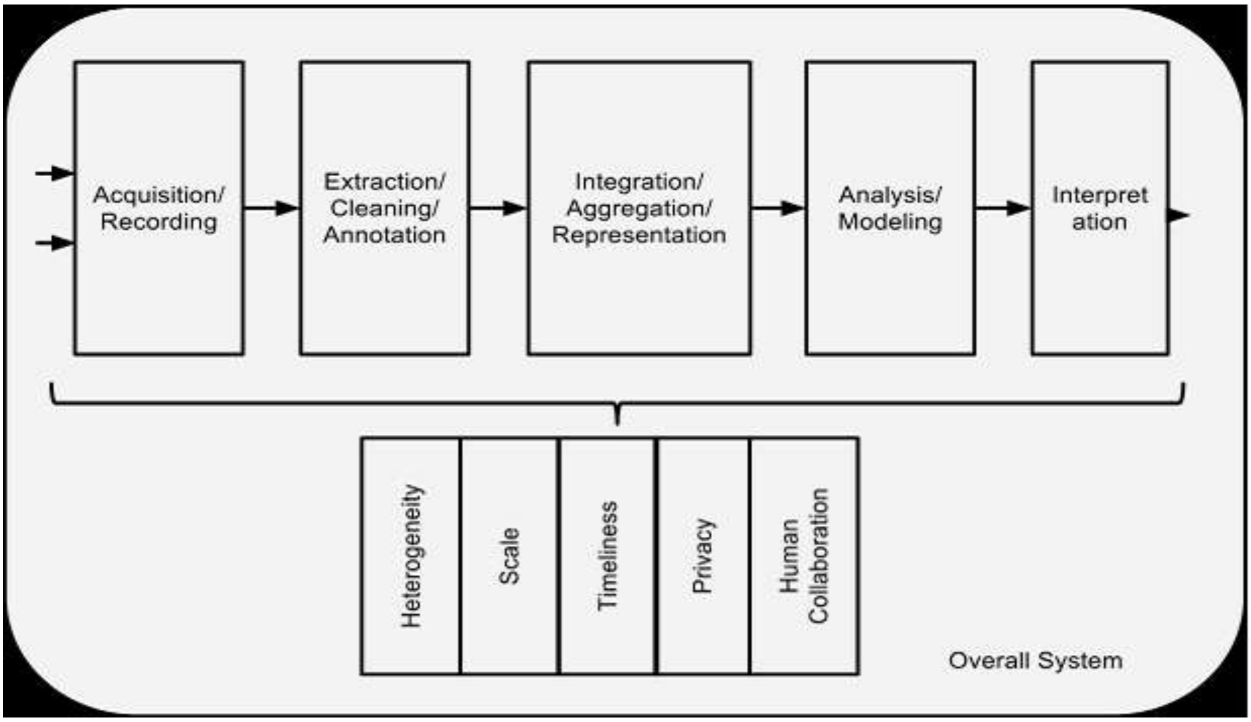

Fig: The BDA pipeline. Major steps in analysis of big data are shown in the flow at top. Below it are big data needs that make these tasks challenging.

Fortunately, existing computational 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 the five stages in the pipeline, then move on to the five cross-cutting challenges, and end with a discussion of the architecture of the overall system that combines all these functions.

II 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 kilometer array telescope, which will produce up to 1 million terabytes of raw data per day. 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. 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.

2.2 Information Extraction and Cleaning

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, 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.

2.3 Data Integration, Aggregation, and Representation

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.

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 Modeling, 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, inter-related 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. 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.

2.5 Interpretation

A decision-maker, provided with the result of analysis, has to interpret the results. This interpretation cannot happen in a vacuum. Usually, it involves examining all the assumptions made and retracing the analysis. Furthermore, as we saw above, there are many possible sources of error: computer systems can have bugs, models almost always have assumptions, and results can be based on erroneous data. For all of these reasons, no responsible user will cede authority to the computer system. Rather she will try to understand, and verify, the results produced by the computer. This is particularly a challenge with Big Data due to its complexity. There are often crucial assumptions behind the data recorded. Analytical pipelines can often involve multiple steps, again with assumptions built in. In short, it is rarely enough to provide just the results. Rather, one must provide supplementary information that explains how each result was derived, and based upon precisely what inputs. Such supplementary information is called the provenance of the (result) data.

III CHALLENGES IN BIG DATA ANALYSIS

Having described the multiple phases in the Big Data analysis pipeline, we now turn to some common challenges that underlie many, and sometimes all, of these phases.

3.1 Heterogeneity and Incompleteness

When humans consume information, a great deal of heterogeneity is comfortably tolerated. However, machine analysis algorithms expect homogeneous data, and cannot understand gradation. In consequence, data must be carefully structured as a first step in data analysis. The three design choices listed have successively less structure and, conversely, successively greater variety. Greater structure is likely to be required by many data analysis systems. However, the less structured design is likely to be more effective for many. However, computer systems work most efficiently if they can store multiple items that are all identical in size and structure.

Even after data cleaning and error correction, some incompleteness and some errors in data are likely to remain. This incompleteness and these errors must be managed during data analysis. Doing this correctly is a challenge. Recent work on managing probabilistic data suggests one way to make progress.

3.2 Scale

The first thing anyone thinks of with Big Data is its size. After all, the word “big” is there in the very name. Managing large and rapidly increasing volumes of data has been a challenging issue for many decades. In the past, this challenge was mitigated by processors getting faster, following Moore’s law, to provide us with the resources needed to cope with increasing volumes of data. But, there is a fundamental shift underway now: data volume is scaling faster than compute resources, and CPU speeds are static.

First, over the last five years the processor technology has made a dramatic shift - rather than processors doubling their clock cycle frequency every 18-24 months, now, due to power constraints, clock speeds have largely stalled and processors are being built with increasing numbers of cores.

The second dramatic shift that is underway is the move towards cloud computing, which now aggregates multiple disparate workloads with varying performance goals (e.g. interactive services demand that the data processing engine return back an answer within a fixed response time cap) into very large clusters. This level of sharing of resources on expensive and large clusters requires new ways of determining how to run and execute data processing jobs so that we can meet the goals of each workload cost-effectively, and to deal with system failures, which occur more frequently as we operate on larger and larger clusters.

A third dramatic shift that is underway is the transformative change of the traditional I/O subsystem. For many decades, hard disk drives (HDDs) were used to store persistent data. HDDs had far slower random IO performance than sequential IO performance, and data processing engines formatted their data and designed their query processing methods to “work around” this limitation.

3.3 Timeliness

The flip side of size is speed. The larger the data set to be processed, the longer it will take to analyze. The design of a system that effectively deals with size is likely also to result in a system that can process a given size of data set faster. However, it is not just this speed that is usually meant when one speaks of Velocity in the context of Big Data. There are many situations in which the result of the analysis is required immediately. Given a large data set, it is often necessary to find elements in it that meet a specified criterion. In the course of data analysis, this sort of search is likely to occur repeatedly. Scanning the entire data set to find suitable elements is obviously impractical. Rather, index structures are created in advance to permit finding qualifying elements quickly.

3.4 Privacy

The privacy of data is another huge concern, and one that increases in the context of Big Data. For electronic health records, there are strict laws governing what can and cannot be done. Managing privacy is effectively both a technical and a sociological problem, which must be addressed jointly from both perspectives to realize the promise of big data.

Consider, for example, data gleaned from location-based services. These new architectures require a user to share his/her location with the service provider, resulting in obvious privacy concerns. Note that hiding the user's identity alone without hiding her location would not properly address these privacy concerns. An attacker or a (potentially malicious) location-based server can infer the identity of the query source from its (subsequent) location information. This is because with location-based services, the location of the user is needed for a successful data access or data collection, while the identity of the user is not necessary.

3.5 Human Collaboration

In spite of the tremendous advances made in computational analysis, there remain many patterns that humans can easily detect but computer algorithms have a hard time finding. Indeed, CAPTCHAs exploit precisely this fact to tell human web users apart from computer programs.

In today's complex world, it often takes multiple experts from different domains to really understand what is going on. A Big Data analysis system must support input from multiple human experts, and shared exploration of results. These multiple experts may be separated in space and time when it is too expensive to assemble an entire team together in one room. The data system has to accept this distributed expert input, and support their collaboration.

IV CONCLUSION

We have entered an era of Big Data. Through better analysis of the large volumes of data that are becoming available. However, many technical challenges described in this paper must be addressed before this potential can be realized fully. The challenges include not just the obvious issues of scale, but also heterogeneity, lack of structure, error-handling, privacy, timeliness, provenance, and visualization, at all stages of the analysis pipeline from data acquisition to result interpretation. These technical challenges are common across a large variety of application domains, and therefore not cost-effective to address in the context of one domain alone. Furthermore, these challenges will require transformative solutions, and will not be addressed naturally by the next generation of industrial products.

In databases, there is considerable work on optimizing individual operations, such as joins. It is well-known that there can be multiple orders of magnitude difference in the cost of two different ways to execute the same query. Fortunately, the user does not have to make this choice – the database system makes it for her. In the case of Big Data, these optimizations may be more complex because not all operations will be I/O intensive as in databases. So standard database optimization techniques cannot directly be used. However, it should be possible to develop new techniques for Big Data operations inspired by database techniques.

The very fact that Big Data analysis typically involves multiple phases highlights a challenge that arises routinely in practice: production systems must run complex analytic pipelines, or workflows, at routine intervals, e.g., hourly or daily. New data must be incrementally accounted for, taking into account the results of prior analysis and pre-existing data. And of course, provenance must be preserved, and must include the phases in the analytic pipeline. Current systems offer little to no support for such Big Data pipelines, and this is in itself a challenging objective.

REFERENCES

[1] The Age of Big Data. Steve Lohr. *New York Times*, Feb 11, 2012. http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html

[2] Kuo, M-H., Sahama, T., Kushniruk, A.W., Borycki, E.M. and Grunwell, D.K. (2014) ‘Health big data analytics: current perspectives, challenges and potential solutions’, *Int. J. Big Data Intelligence*, Vol. 1, Nos. 1/2, pp.114–126.

[3] *Nature* (2012) ‘Seven days – the news in brief’, Vol. 484, pp.10–11.

[4] Agency for Healthcare Research and Quality, *What Is Comparative Effectiveness Research* [online] http://effectivehealthcare.ahrq.gov/index.cfm/what-iscomparative-effectiveness-research1/ (accessed 2 November 2013).

[5] Aggarwal, C. and Wang, H. (2010) ‘Managing and mining graph data’, *Series: Advances in Database Systems*, Vol. 40, Springer, ISBN 978-1-4419-6045-0.

[6] Aggarwal, C.C. and Yu, P.S. (2008) *Privacy-Preserving Data Mining- Models and Algorithms*, Springer, ISBN 978-0-387-70991-8.

[7] Agrawal, D. et al. (2012) *Challenges and Opportunities with Big Data*, Big Data White Paper-Computing Research Association [online] http://cra.org/ccc/docs/init/bigdatawhitepaper.pdf (accessed 5 November 2013).

[8] http://www.cra.org/ccc/files/docs/init/bigdatawhitepaper.pdf

[9] http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2819144/