INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X

(Volume2, Issue5)

Available online at: www.ljariit.com

Big Data: Challenges and Opportunities

Meena Rani*
meenarani912@gmail.com

Ridhi Jindal
Ridhijindal136@gmail.com

Abstract— Big Data has the potential to revolutionize not just research, but also education. There are few defines phases through which the data is to be processed. There are challenges like Heterogeneity and Incompleteness, scale, timeliness, privacy and human consideration. Through better analysis of the large volumes of data that are becoming available, there is the potential for making faster advances in many scientific disciplines and improving the profitability and success of many enterprises.

Keywords— Big Data, Scale, hetroginity,Visualization.

I. INTRODUCTION

Scientific research has been revolutionized by Big Data. The Sloan Digital Sky Survey 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 curation 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.

II. BASIC MODEL

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/modeling 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.

Phases in the Processing Pipeline:

1. Data Acquisition and Recording.
2. Information Extraction and Cleaning.
3. Data Integration, Aggregation, and Representation.
4. Query Processing, Data Modeling, and Analysis.
5. Interpretation.

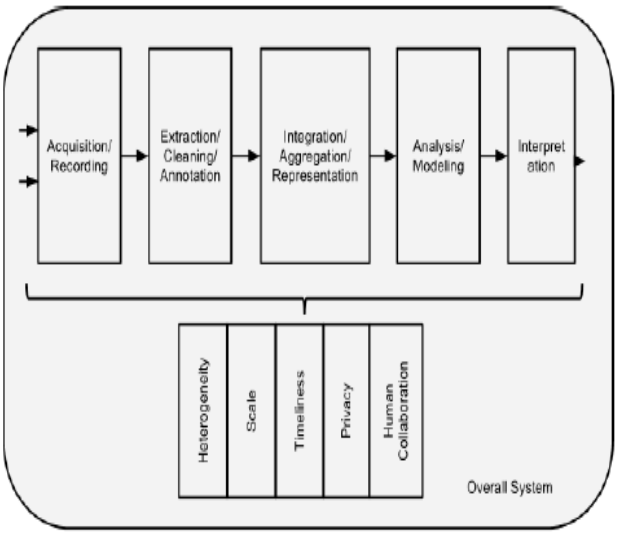

Figure 1: The Big Data Analysis Pipeline. Major steps in analysis of big data are shown in the flow at top. Below it are big data needs that make these tasks challenging.

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.

1. Heterogeneity and Incompleteness

When humans consume information, a great deal of heterogeneity is comfortably tolerated. In fact, the nuance and richness of natural language can provide valuable depth. However, machine analysis algorithms expect homogeneous data, and cannot understand nuance. In consequence, data must be carefully structured as a first step in (or prior to) data analysis. Consider, for example, a patient who has multiple medical procedures at a hospital. We could create one record per medical procedure or laboratory test, one record for the entire hospital stay, or one record for all lifetime hospital interactions of this patient. With anything other than the first design, the number of medical procedures and lab tests per record would be different for each patient. The three design choices listed have successively less structure and, conversely, successively greater variety. Greater structure is likely to be required by many (traditional) data analysis systems. However, the less structured design is likely to be more effective for many purposes – for example questions relating to disease progression over time will require an expensive join operation with the first two designs, but can be avoided with the latter. However, computer systems work most efficiently if they can store multiple items that are all identical in size and structure. Efficient representation, access, and analysis of semi-structured data require further work.

2. Scale

Of course, 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.

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. For example, if a fraudulent credit card transaction is suspected, it should ideally be flagged before the transaction is completed – potentially preventing the transaction from taking place at all. Obviously, a full analysis of a user's purchase history is not likely to be feasible in real-time. Rather, we need to develop partial results in advance so that a small amount of incremental computation with new data can be used to arrive at a quick determination.

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. For other data, regulations, particularly in the US, are less forceful. However, there is great public fear regarding the inappropriate use of personal data, particularly through linking of data from multiple sources. 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.

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. Ideally, analytics for Big Data will not be all computational – rather it will be designed explicitly to have a human in the loop. The new sub-field of visual analytics is attempting to do this, at least with respect to the modeling and analysis phase in the pipeline. There is similar value to human input at all stages of the analysis pipeline.

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. SYSTEM ARCHITECTURE

With Big Data, the use of separate systems in this fashion becomes prohibitively expensive given the large size of the data sets. The expense is due not only to the cost of the systems themselves, but also the time to load the data into multiple systems. In consequence, Big Data has made it necessary to run heterogeneous workloads on a single infrastructure that is sufficiently flexible to handle all these workloads. The challenge here is not to build a system that is ideally suited for all processing tasks. Instead, the need is for the underlying system architecture to be flexible enough that the components built on top of it for expressing the various kinds of processing tasks can tune it to efficiently run these different workloads.

If users are to compose and build complex analytical pipelines over Big Data, it is essential that they have appropriate high-level primitives to specify their needs in such flexible systems. The Map-Reduce framework has been tremendously valuable, but is only a first step.

Even declarative languages that exploit it, such as Pig Latin, are at a rather low level when it comes to complex analysis tasks. Similar declarative specifications are required at higher levels to meet the programmability and composition needs of these analysis pipelines. Besides the basic technical need, there is a strong business imperative as well. Businesses typically will outsource Big Data processing, or many aspects of it. Declarative specifications are required to enable technically meaningful service level agreements, since the point of the out-sourcing is to specify precisely what task will be performed without going into details of how to do it.

V. CONCLUSION

We have entered an era of Big Data. Through better analysis of the large volumes of data that are becoming available, there is the potential for making faster advances in many scientific disciplines and improving the profitability and success of many enterprises. 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.

REFERENCES

[1] Advancing Discovery in Science and Engineering. Computing Community Consortium. Spring 2011.
[2] Advancing Personalized Education. Computing Community Consortium. Spring 2011.
[3] Smart Health and Wellbeing. Computing Community Consortium. Spring 2011.
[4] A Sustainable Future. Computing Community Consortium. Summer 2011.
[5] Getting Beneath the Veil of Effective Schools: Evidence from New York City. Will Dobbie, Roland G. Fryer, Jr. NBER Working Paper No. 17632. Issued Dec. 2011.

[6] Drowning in numbers -- Digital data will flood the planet—and help us understand it better. The Economist, Nov18,2011.http://www.economist.com/blogs/dailychart/2011/11/big-data-0
[7] Using Data for Systemic Financial Risk Management. Mark Flood, H V Jagadish, Albert Kyle, Frank Olken, and Louiqa Raschid. Proc. Fifth Biennial Conf. Innovative Data Systems Research, Jan. 2011.
[8] Pattern-Based Strategy: Getting Value from Big Data. Gartner Group press release. July 2011. Available at http://www.gartner.com/it/page.jsp?id=1731916
[9] Understanding individual human mobility patterns. Marta C. González, César A. Hidalgo, and Albert-László Barabási. Nature 453, 779-782 (5 June 2008)
[10] Computational Social Science. David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer,Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. Science 6 February 2009: 323 (5915), 721-723.
[11] Big data: The next frontier for innovation, competition, and productivity. James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung Byers. McKinsey Global Institute. May 2011.
[12] Materials Genome Initiative for Global Competitiveness. National Science and Technology Council. June 2011.
[13] Folowing the Breadcrumbs to Big Data Gold. Yuki Noguchi. National Public Radio, Nov. 29, 2011. http://www.npr.org/2011/11/29/142521910/the-digital-breadcrumbs-that-lead-to-big-data
[14] The Search for Analysts to Make Sense of Big Data. Yuki Noguchi. National Public Radio, Nov. 30, 2011. http://www.npr.org/2011/11/30/142893065/the-search-for-analysts-to-make-sense-of-big-data
[15] 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 15

