Challenges and Opportunities with Big Data

Alexandros Labrinidis
University of Pittsburgh
Pittsburgh, PA, USA
labrinid@cs.pitt.edu

H. V. Jagadish
University of Michigan
Ann Arbor, MI, USA
jag@umich.edu

ABSTRACT

The promise of data-driven decision-making is now being recognized broadly, and there is growing enthusiasm for the notion of "Big Data," including the recent announcement from the White House about new funding initiatives across different agencies, that target research for Big Data. While the promise of Big Data is real – for example, it is estimated that Google alone contributed 54 billion dollars to the US economy in 2009 – there is no clear consensus on what is Big Data. In fact, there have been many controversial statements about Big Data, such as "Size is the only thing that matters." In this panel we will try to explore the controversies and debunk the myths surrounding Big Data.

1. INTRODUCTION

It is hard to avoid mention of Big Data anywhere we turn today. There is broad recognition of the value of data, and products obtained through analyzing it [1]. Popular news media now appreciates the value of Big Data as evidenced by coverage in the Economist [2, 3], the New York Times [5], and National Public Radio [7, 8]. Industry is abuzz with the promise of Big Data [6]. Government agencies have recently announced significant programs towards addressing challenges of Big Data[1]. Yet, many have a very narrow interpretation of what that means, and we lose track of the fact that there are multiple steps to the data analysis pipeline, whether the data are big or small. At each step, there is work to be done, and there are challenges with Big Data.

The first step is data acquisition. Some data sources, such as sensor networks, can produce staggering amounts of raw data. 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, in considering news reports, is it enough to retain only those that mention the name of a company of interest? Do we need the full report, or just a snippet around the mentioned name? The second big challenge is to automatically generate the right metadata to describe what data is recorded and how it is recorded and measured. This metadata is likely to be crucial to downstream analysis. For example, we may need to know the source for each report if we wish to examine duplicates.

Frequently, the information collected will not be in a format ready for analysis. The second step is an information extraction process that pulls out the required information from the underlying sources and expresses it in a structured form suitable for analysis. A news report will get reduced to a concrete structure, such as a set of tuples, or even a single class label, to facilitate analysis. Furthermore, we are used to thinking of Big Data as always telling us the truth, but this is actually far from reality. We have to deal with erroneous data: some news reports are inaccurate.

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. 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.

Mining requires integrated, cleaned, trustworthy, and efficiently accessible data, declarative query and mining interfaces, scalable mining algorithms, and Big Data computing environments. 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.

Having the ability to analyze Big Data is of limited value if users cannot understand the analysis. Ultimately, a decision-maker, provided with the result of analysis, has to interpret these results. Usually, this 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, users will try to understand, and verify, the results produced by the computer. The computer system must make it easy for her to do so by providing supplementary information that explains how each result was derived, and based upon precisely what inputs.

In short, there is a multi-step pipeline required to extract value from data. Heterogeneity, incompleteness, scale, timeliness, pri-

[1] http://www.whitehouse.gov/blog/2012/03/29/big-data-big-deal

Proceedings of the VLDB Endowment, Vol. 5, No. 12

vacy and process complexity give rise to challenges at all phases of the pipeline. Furthermore, this pipeline is not a simple linear flow – rather there are frequent loops back as downstream steps suggest changes to upstream steps. There is more than enough here that we in the database research community can work on [4].

2. PANEL GOALS

The panel has multiple goals. First of all, to identify if/why Big Data is different from past Very Large DataBase techniques and what are the most challenging aspects of Big Data. Secondly, to determine how can the (data management) industry and academia collaborate towards solving Big Data challenges. Finally, to consider the role of the data management community within the Big Data solutions "ecosystem."

In order to address these goals, the panel will discuss the validity of the following claims:

- Big Data is the same as scalable analytics.
- Big Data problems are primarily at the application side.
- Big Data problems are primarily at the systems level.
- Big Data requires a cloud-based platform.
- The Data Management community is in danger of missing the Big Data train.
- It is not possible to conduct Big Data research effectively without collaborating with people outside the data management community.
- All the Big Data problems can be reduced to Map/Reduce problems.
- The bulk of Big Data challenges are being addressed by industry.
- The bulk of Big Data challenges are at the implementation level.
- Size is the only thing that matters (for Big Data).

3. PANEL COMPOSITION

3.1 Moderators

Alexandros Labrinidis is an associate professor at the Department of Computer Science of the University of Pittsburgh and co-director of the Advanced Data Management Technologies Lab. He is also an adjunct associate professor at Carnegie Mellon University (CS Dept). He is currently the Secretary/Treasurer for ACM SIGMOD, and has served as the Editor of SIGMOD Record. He is the recipient of an NSF CAREER award in 2008.

H. V. Jagadish is Bernard A Galler Collegiate Professor of Electrical Engineering and Computer Science, and Director of the Software Systems Research Laboratory, at the University of Michigan in Ann Arbor. He is a fellow of the ACM, serves on the board of the Computing Research Association, and is Founding Editor-in-Chief of the Proceedings of the VLDB Endowment (since 2008).

3.2 Panelists

Susan Davidson, Weiss Professor and Chair, Department of Computer and Information Science, Founding Co-Director Center for Bioinformatics, and previously Deputy Dean, School of Engineering and Applied Sciences, all at the University of Pennsylvania. Also co-Founder, Greater Philadelphia Bioinformatics Alliance.

Johannes Gehrke, Tisch University Professor, Department of Computer Sciences, Cornell University. He has received the Cornell University Provost's Award for Distinguished Scholarship, a Humboldt Research Award from the Alexander von Humboldt Foundation, the 2011 IEEE Computer Society Technical Achievement Award, and the 2011 Blavatnik Award for Young Scientists from the New York Academy of Sciences.

Nick Koudas, Professor, Department of Computer Science, University of Toronto and President & Co-founder, Sysomos. He was named 2011 Inventor of the year by the University of Toronto.

Raghu Ramakrishnan, Technical Fellow; CTO, Information Services; Director, Cloud and Information Services Lab, Microsoft. He received the ACM SIGKDD Innovation Award in 2008 and the ACM SIGMOD Contributions Award in 1999. He was elected Fellow of the ACM in 2001 and Fellow of the IEEE in 2008.

4. REFERENCES

[1] Big Data. Nature (http://www.nature.com/news/specials/bigdata/index.html), Sep 2008.

[2] Data, data everywhere. The Economist (http://www.economist.com/node/15557443), Feb 2010.

[3] Drowning in numbers – Digital data will flood the planet—and help us understand it better. The Economist (http://www.economist.com/blogs/dailychart/2011/11/big-data-0), Nov 2011.

[4] D. Agrawal, P. Bernstein, E. Bertino, S. Davidson, U. Dayal, M. Franklin, J. Gehrke, L. Haas, A. Halevy, J. Han, H. V. Jagadish, A. Labrinidis, S. Madden, Y. Papakonstantinou, J. M. Patel, R. Ramakrishnan, K. Ross, C. Shahabi, D. Suciu, S. Vaithyanathan, and J. Widom. Challenges and Opportunities with Big Data – A community white paper developed by leading researchers across the United States. http://cra.org/ccc/docs/init/bigdatawhitepaper.pdf, Mar 2012.

[5] S. Lohr. The age of big data. New York Times (http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html), Feb 2012.

[6] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers. Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute, May 2011.

[7] Y. Noguchi. Following Digital Breadcrumbs to Big Data Gold. National Public Radio (http://www.npr.org/2011/11/29/142521910/the-digital-breadcrumbs-that-lead-to-big-data), Nov 2011.

[8] Y. Noguchi. The Search for Analysts to Make Sense of Big Data. National Public Radio (http://www.npr.org/2011/11/30/142893065/the-search-for-analysts-to-make-sense-of-big-data), Nov 2011.