Rich Data: Risks, Issues, Controversies & Hype

Osmar R. Zaïane
Department of Computing Science
University of Alberta, Canada
zaiane@cs.ualberta.ca

Abstract

Big data technology is being adopted in industry and government at a record rate. A large number of enterprises believe big data analytics will redefine the competitive landscape of their industries within the next few years. Adoption is now perceived as a matter of survival. The unprecedented accumulation of data in almost all industries and government is unquestionable, but is the extravagant promotion of the technology justifiable? Is the technology ready to deliver on the promises? And is the fear driving the technology adoption reasonable? We will try to shed some light on the current state of rich data.

1 Introduction

The continuously increasing deluge of complex data we experience today is undeniable. If there is a hype about big data it is not about whether it is already upon us, but possibly on the expectations about what we can currently attain from its analytics. The buzzword Big Data is unfortunately a misnomer. It is an inaccurate term since it is misleading to understand the real significance of the idiom, even for specialists in information technology. Most focus on the qualifier "Big" to emphasize solely the size and miss the most important nature of the medium, the complexity of the data. Big Data refers to the massive amounts of complex data that are difficult to manipulate and understand using traditional processing methods. The complexity is not only due to the size but many other factors we highlight later. Thus, we advocate the designation Rich Data. What added to the confusion is the issue of the journal Nature on Big Data (Doctorow, 2008) that mainly centered on the issue of size. Big data was originally used rhetorically (Anderson, 2008) indicating that big is a fast moving target when it comes to data. What we considered large is not anymore, and what we consider huge today will not be soon. For the originators of the term (Anderson, 2008) Big Data typically meant applying tools such as Machine Learning to vast data beyond that captured in standard databases. Examples of such data include web browsing trails, social media, sensor data, surveillance data, etc. Based on this definition, big data is today ubiquitous and inescapable. This definition also hints to the moving target again; big data refers to rich complex data for which existing methods for storage, indexing, processing and analyzes are inadequate and new methods are required. As soon as solutions are found, big data is again something else for which methods have yet to be devised. The clever IBM marketing team has presented Big Data in terms of four dimensions now commonly known as the 4 Vs: Volume, Velocity, Variety, and Veracity (IBM, 2015). Today, most see big data as a mix of structured, semi-structured, and unstructured data, which typically breaks barriers for traditional relational database storage and breaks the limits of indexing by "rows". Hence the emergence of No-SQL and NewSQL data stores using a simple data model based on key-value pairs (Grolinger, 2013). This data also typically requires intensive pre-processing before each query to extract "some structure", particularly when it comes to text, and entails massively parallel and distributed computing with Map-Reduce type operations; to the point that Hadoop, an open source framework implementing the Map-Reduce processing, is becoming synonymous with Big Data (White, 2012). In reality, there is no standard recipe or architecture for big data. Big Rich Data is when we have complex data coming from disparate data sources that require integration to extract real value. Each problem is unique, hence the need for data scientists, who are not only data analysts but specialists contemplating holistic solutions considering infrastructures for data storage and management, as well as methods for aggregating,

analyzing and visualizing data and patterns. Data scientists do not work in isolation but in teams bringing together different skills in data analytics.

2 The Famous Vs of Big Data

IBM is credited for introducing the dimensions of Big Data. They were initially three (Volume, Velocity and Variety) and later augmented with Veracity (IBM, 2015). In fact, other Vs have been proposed later by other data science experts. We introduce herein 7 Vs.

Volume: refers to the size of the data which is typically very large. We are indeed awash with data, be it scientific data, data generated from activities on the web, acquired from sensors or collected from social media. We have an enormous volume of data at our disposal and are witnessing an exponential growth. However, not all problems with large volume of data are big data problems, and not all big data problems are concerned with very large data.

Velocity: is concerned with the speed of data creation and the speed of change. Sensors continuously transmit their measures; trades are done in milliseconds; credit card transactions are conducted world-wide uninterruptedly; social media messages go constantly viral in minutes. This velocity of the data is equated to a firehose of data from which we can read the data only once and having to analyze it while it is being generated. Velocity for rich data refers also to the speed of required analysis. Analysis and reporting of the results are also constraint with time.

Variety: refers to the different types of data we can now use, but more importantly refers to the vast array of data sources at our disposal. In the past, applications mainly exploited numerical and categorical data stored in relational tables, called structured data; with Rich Data applications we need to harness differed types of data including, images, video sequences, voice, time series, text messages from social media, and last but not least the relationships between data objects such as in social networks. Variety comes also from the availability of myriad independent data sources sometimes even from the public domain, such as open-data or from the Web. Acquiring and integrating additional data to the available one enhances the insights that can be obtained from the original data.

Veracity: Available data is often uncertain, particularly when acquired from sources over which we do not have control, such as social media. Veracity refers to ascertaining the accuracy of the analysis results or understanding of the discovered information when uncertainty prevails in the source data. The volume of data often makes up for the lack of quality or accuracy, but models that provide probabilistic results are preferred to measure some trust in the results.

Value: refers to the capacity to transform data into value, and more often the value is in the integration of data from different autonomous sources. The power of big data is to leverage additional independent data sources to better extract actionable knowledge and new information from an original dataset to bring more value in a decision making process.

Visualization: encompasses the reporting of the results of the analysis and effectively communicating actionable knowledge to decision makers. Visualization is the art of coalescing complex information into one 2D or 3D possibly interactive image. It is the essential lens through which one can see and understand the patterns and the relationships in the data.

Vulnerability: pertains to the privacy of the data that could be jeopardized. This is often the forgotten V. Even when dealing with anonymized data, when combining with additional data from other separate sources, the integration can reveal previously undisclosed information and thus expose private evidence. Data anonymization is typically attacked and compromised by combining data sources, which is the essence of big data. Privacy preserving techniques need to be intrinsic to big data analytics.

3 The Value is in Data Integration

A concrete example can illustrate the spirit of big data analytics. In 2006, wanting to improve on its Cinematch recommender system, the Netflix company launched a $1M challenge to whom would improve the results of their algorithm by at least 10%. The competition was clear on not to use other data sources but the 100M ratings in a sparse matrix of 500k users and 17k movies. It took about 3 years to win the prize with an improvement equivalent to 1/10 of a star. The solution was too convoluted for Netflix to implement. It was not the complexity of the solution the main reason for dropping it, but the realization that using additional information such as the Internet Movie Database (IMDB) with information on actors, directors, etc. and their relationships as well as sentiment in reviews could provide additional value to the ratings in the matrix to deliver better results for a recommender system with a more powerful

predictive model (Amatriain, 2013). The lesson learned is that one should always exploit all obtainable data, not just the data available at hand.

4 The Pitfalls & Challenges of Big Data

There is hype when the rate of adoption out paces the ordinary evolution of the technology and to avoid a quick disillusionment towards the technology one must manage to balance between the expectations and the promises. This same imbalance led to the disappointment toward Artificial Intelligence and its relinquishment by the major funders in the 1970s and again in the late 1980s, periods known as the winters of AI. It is debated whether Big Data would know such winter with a serious dwindling of the investment. The value of data is commonly agreed upon, yet very few know how to profit of this data for competitive advantage. Where big data has undeniably seen success is in consumer behaviour prediction but the very quick adoption is touching all industries and government. Many have invested significant amounts of money in the technology mainly by fear of missing the train of opportunity, but the interest can fade since many are failing to realize and operationalize the value that lies in big data and the voluminous investment that comes with it. For the adoption to endure and to drive more innovation, the community must be more mindful of the technology and cognizant of the pitfalls and challenges. We highlight some herein.

Few years ago an authoritative report created a stir in the industry. The McKinsey Report asserted that in the US alone there will be a shortage by 2018 of up to 190,000 data scientists (Manyika, 2011). This led the Harvard Business Review to state data scientist as being the "Sexiest Job" in this century (Davenport, 2012). Training more data scientists with deep analytical skills is becoming a necessity. Meanwhile, with the current void, we have many that deceptively claim knowledge and skills in data science which could contribute to the disillusionment. The McKinsey Report also stressed the necessity to educate managers in the know-how to use the analysis of big data to make effective decisions. Educating managers gives them the opportunity to leverage the skills of their data science team and surely take advantage of big data analytics.

Another important downside is one of the least insisted upon V of big data: Veracity. The voluminous size is a curse for big data as with vast data, patterns can happen by chance but these patterns may have no predictive power. Like with statistics, facts in data are vulnerable to misuse and with this pliability of data one can make it mean anything. Data per se does not create meaning but data analysts make it express the hidden information and bring forth the crucial interpretation. As Susan Etlinger articulated it: "Critical Thinking is the killer app for Big Data" (Etlinger, 2014). Hence the need for the data context, known as metadata. Metadata, describing the data, should be created at the source, should journey with the data, managed with the data, exploited during analysis, and used for interpretation. A pillar component of big data is data fusion, but integrating data cannot be safely accomplished without using metadata. Metadata is also paramount for the interpretation of patterns as well as visualizing and clarifying analysis results. Likewise, visualization is still a neglected limitation while it is of paramount importance in any complete data mining process as it conveys the final discoveries (Fayyad, 2001). Visualization, the visual reporting of discoveries from data, is not a real science but an art; the art of conveying patterns from a high dimensional space in 2D representation, possibly interactively, without losing information while highlighting the essential and actionable. One typical mistake is not to work with skilled artists and trained communication specialists who have different perspectives and think outside the box to produce such required visualizations.

Big Data carries challenges for the scientific community. The challenges are numerous which represent huge opportunities for research and innovation. The first challenge is obviously the scale. The trend is going towards collecting even more data and the advent of the Internet of Things will only be a multiplier (Greengard, 2015). The challenge is not only building the required infrastructure to store and manage the data but also analyzing it efficiently and obtain the valuable insights. The popular MapReduce concept has become the generic programming model used to store and process large scale datasets on commodity hardware clusters. However, not all problems are "map-reducible". New initiatives for massive distributed computing, such as the Spark framework (Karau, 2013), are already being introduced. Another defiant problem is due to data heterogeneity from various sources and data inconsistencies between sources. While data integration and record linking has a long tradition in the database research community, it is still in its infancy when

it comes to rich data and its complexities. Combining different data sources brings additional challenges such as data incompleteness and uncertainty, which again highlight the importance of Veracity. Last but not least, combining data sources also creates a possible confrontation with data privacy. Truly privacy-preserving data mining techniques can compromise data utility (Wu, 2013). Anonymization approaches add perturbations to generate altered versions of the data with additional uncertainties. It remains that data sharing in big data raises many security and privacy concerns. Another overlooked challenge is the one due to the data dimensionality explosion (Wu, 2014). Big Data is also concerned with large dynamic and growing complex data. In this active data, not only are we faced with high and diverse dimensionality issues, but the dimensions keep changing with new additions, disappearances and modifications. The ultimate challenge is automation of the big data analytics such as with autonomous vehicles. There is a trend towards analysis for non-data scientists; creating generic mechanized systems taking disparate data sources as input and producing reports with a push of a button.

5 Conclusion

We are constantly bombarded by stories about how much data there is in the world and how traditional solutions are too slow, too small, or too expensive to use for such large data, but when it comes to Rich Data and the challenges of interpreting it, size is not everything. There is also speed at which it is created and the variety of it and its complexity of types and sources. Because Big Data could improve decision making in myriad fields from business to medicine allowing decisions to be based on data and data analysis, large corporations have adopted Big Data in their decision making, predominantly in marketing and customer behavior analysis. Big Data is only getting worse in terms of volume, speed, availability of sources and complexity, and most sectors of the economy are data-driven decision making. Therefore, big data is not just a buzzword anymore, but to avoid a hype we must manage the realistic expectations. Otherwise, people may be quickly disappointed by not getting what is promised and what is currently possible. A common gaffe is to focus on infrastructure, yet a holistic solution is required: data linking is part of the solution, new hardware is part of the solution, and new algorithms are part of the solution. The key is to deploy all means to be able to exploit all the data that is obtainable to enhance insights and possible actions.

Reference

Chris Anderson, 2008, The Petabyte Age: Because More Isn't Just More — More Is Different, Wired Magazine, Issue 16.07

Xavier Amatriain, 2013, Mining large streams of user data for personalized recommendations. ACM SIGKDD Explorations Newsletter, 14(2):3748,

Randal E. Bryant, Randy H. Katz, Edward D. Lazowska, 2008, Big-Data Computing: Creating revolutionary breakthroughs in commerce, science, and society, Computing Community Consortium.

Thomas Davenport, D.J. Patil, 2012, Data Scientist: The sexiest Job of the 21st Century, Harvard Business Review

Cory Doctorow, 2008, Welcome to the petacentre, Big Data Special issue, Nature 455, 1

Susan Etlinger, 2014, Critical Thinking: The Killer App for Big Data, TED Talk, https://www.ted.com/talks/susan_etlinger_what_do_we_do_with_all_this_big_data

Usama Fayyad, Georges Grinstein, Andreas Wierse, 2001, Information Visualization in Data Mining and Knowledge Discovery, Morgan Kaufmann.

Samuel Greengard, 2015, The Internet of Things, MIT Press

Katarina Grolinger, Wilson Higashino, Abhinav Tiwari, Miriam Capretz, 2013, Data management in cloud environments: NoSQL and NewSQL data stores, Journal of Cloud Computing: Advances, Systems and Applications, 2:22, Springer.

IBM, 2015, The Four V's of Big Data, http://www.ibmbigdatahub.com/infographic/four-vs-big-data

Holden Karau, 2013, Fast Data Processing with Spark, Packt Publishing

James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, Angela Hung Byers, 2011, Big data: The next frontier for innovation, competition, and productivity, McKinsey Global Institute

Tom White, 2012, Hadoop: The definitive guide, O'Reilly Media, Inc.

Lengdong Wu, Hua He, Osmar Zaiane, 2013, Utility Enhancement for Privacy Preserving Health Data Publishing, International Conference on Advanced Data Mining and Applications

Xindong Wu, Xingquan Zhu, Gong-Qing Wu, Wei Ding, 2014, Data Mining with Big Data, IEEE Transactions on Knowledge & Data Engineering, vol.26, Issue 1.