Big Data Research 2 (2015) 49–52

Contents lists available at ScienceDirect

Big Data Research

www.elsevier.com/locate/bdr

Big Data and Science: Myths and Reality ☆

H.V. Jagadish

University of Michigan, United States

ARTICLE INFO

Article history:
Received 13 November 2014
Accepted 11 January 2015
Available online 23 February 2015

ABSTRACT

As Big Data inexorably draws attention from every segment of society, it has also suffered from many characterizations that are incorrect. This article explores a few of the more common myths about Big Data, and exposes the underlying truths.

"Big Data" now impacts nearly every aspect of our modern society, including business, government, health care, and research in almost every discipline: life sciences, engineering, natural sciences, art & humanities. As it has drawn much attention, and become economically important, there are many who have preferred angles on the interpretation of Big Data. At the same time, as many have been exposed to the term with little prior knowledge of computing or technology, they are easily swayed by the "experts." In consequence, there has been a rush to use the term Big Data in ways that are inappropriate but self-serving. In many cases, these erroneous interpretations have then been taken up and amplified by others, including even technically sophisticated people. In this article, I discuss some of the more common myths.

1. Big Data Myth 1: Size is all that matters

The very word "Big" indicates size. It is also the case that measures of size are very easily conveyed. We have all heard statements about how high a stack of phonebooks is required to store the data that is easily kept on one disk drive. So it is not surprising that for many lay people, Big Data is all about size.

One would think that technical people would know better. Unfortunately, size also lends itself to easy measurement. It is straightforward to count up the number of bytes in some data store, and equally easy to plot a sequence of such measurements on a chart showing exponential growth. In fact, such charts have become so common that even many lay people get the concept. What this leads to, among other things, is serious people apologetically saying that they only have a few hundred gigabytes of data and so are not sure that they really have a Big Data problem. This is sad, because we are putting off so many people we ought to be able to help.

In spite of the points made above, I believe that better sense would have prevailed in our understanding of Big Data if it were not for the economic imperatives of the IT industry. We have today a huge ecosystem of Big Data systems. These systems are, for the most part, innovative: collectively, they constitute a whole new paradigm of scaling. There are many who have problems that require this scale and are amenable to these new architectures. These facts have led to the creation of a new industry segment and benefitted many, all of which is good. But the tremendous progress made in this space has also sucked the Oxygen out of the air for everything else, as it were. Industry wants to talk about volume, for economic reasons. And money speaks.

Several years ago, the Gartner group noticed this undue attention focused on size, and proposed the now famous "3Vs" of Big Data [5]. IBM then pushed for adding a 4th V [6], and this has been accepted by most. So, theoretically, most technical people will tell you that Big Data raises issues of Volume, Velocity, Variety, and Veracity (or at least the first three of these). But then they will immediately go on to discuss how many Petabytes there are in some problem.

I have discussed above, why Volume (or size) gets undue attention. Let me turn now to why I think Variety and Veracity do not get the attention they deserve. One major reason for this lack of attention is that there is no well-accepted measure for either. If there is no measure, it is hard to track progress. If I have a company and develop an innovative system that can handle a slightly larger volume than the competition, I can show this off with measurements against some benchmark. If I am an academic and develop an algorithm that scales better than the competition, I know exactly how to compare my algorithm against the competition and persuade skeptical reviewers. In contrast, consider variety. If I have a product that makes handling variety a little easier, what technical claim can I make that doesn't sound like marketing hype? If I write a paper about a data model that is better at handling variety than

☆ This article belongs to Visions on Big Data.

E-mail address: jag@umich.edu.
URL: http://www.eecs.umich.edu/~jag.

http://dx.doi.org/10.1016/j.bdr.2015.01.005
2214-5796/

the current state of the art, I have to think very hard about how I will compare against the competition and establish the goodness of my idea. Progress is hard in things you cannot measure, in both industry and academia. Variety may be the hardest of the 4Vs to address, but it is the one that people are least motivated to speak about.

Veracity suffers from most of the same problems as Variety. Under very simplistic models, we can at least begin to measure some things, establish some probabilities and some distributions, and so forth. But everyone recognizes that these measures are based on unrealistically simple models: for instance ones that assume independence when we know that is not true. Therefore, measures are taken with a grain of salt, and Veracity is scarcely easier to address than Variety.

To conclude, Volume and Velocity are indeed challenging. But Variety and Veracity are far more challenging. It is time we focused the conversation around Big Data appropriately.

2. Big Data Myth 2: The central challenge with Big Data is that of devising new computing architectures and algorithms

Even if we are consciously thinking about Big Data in terms of the 4Vs, we immediately have a question of determining what the thresholds are to call something "Big". For Variety and Veracity we know this is not even an answerable question, because we do not have measures in the first place. So let us just consider Volume and Velocity. The threshold, for some people, is at the limit of what we know how to handle. Obviously, this is a moving target. But it has the advantage of being inspirational. The fatal (in my opinion) drawback is that it limits the size of the market to 1: there is only one largest deployment in the world at any time (barring ties). Increasing the size of this deployment is definitely a worthwhile challenge, but not one that an entire industry can be built around and an entire academic field developed.

The threshold, in some definitions, then becomes fixed, based on the dominant architecture at some point in time, say 2010. So a data set qualifies, in terms of Volume, as Big Data if it is larger than can be handled using the "standard" architectures in use at the beginning of the Big Data era. With the ever-growing popularity of Map-Reduce style computation, and the plethora of systems and tools in "the Big Data eco-system," we then have a definition that is specific, even if it is both circular and self-serving: a Big Data problem is one that is best addressed using elements drawn from the Big Data "toolbox." This definition is specific because there is general agreement about what tools are in the Big Data toolbox: most tool producers self categorize themselves appropriately. The definition is circular because it really does not define what goes into the toolbox. If we did not have an explicit listing, we would be defining a Big Data tool as a software system that addresses at least some aspects of a Big Data problem, or some such similar statement. The definition is self-serving because it anoints a set of tools and a style of system architecture as "the solution" to the Big Data problem. This definition is wrong because almost everything in the Big Data toolbox is focused on Volume (frequently in conjunction with Velocity), with very little consideration given to Variety and Veracity challenges. I believe that the cloud, and what is today considered the "Big Data Ecosystem," has its place in the constellation of relevant technologies, but is neither a complete solution in itself nor a required piece of every solution.

My own threshold for Big Data is more (along any of the 4 axes) than we know how to handle *in context*. The scientist (or manager) faces a Big Data program when she has too much data to be able to process using the spreadsheet program she knows. The solution in this case may be as simple as moving to a database. But even such an apparently simple transition can have many hidden issues: the spreadsheet's current design may not be suitable for a relational table (for example, a new column may be added every month), there may be interdependencies with other components of some complex workflow, and so on. Identifying and eliminating such barriers is legitimate Big Data work. See, for example, the National Academies report on "Frontiers in Massive Data Analysis" [4].

It is also worth noting that we can buy bigger systems, more machines, faster CPU, and larger disks. But human ability does not scale! Moreover, the sizes that become challenging for humans are often very small for computers. For example, consider a graph with just 40 nodes and 200 edges. Try plotting it on screen with your favorite graph-drawing program and then look for patterns. Even such a small graph is likely to be at the limit of what we can manage with technology today. Big Data poses huge challenges for human interaction. Many of the most interesting problems in the Big Data space deal with facilitating this human interaction.

3. Big Data Myth 3: Analytics is the central problem with Big Data

It is completely understandable that many lay people picture a Big Data System as a magic piece of software that takes Big Data as input and produces deep insights as output. Unfortunately, this misperception suits many companies, and even some academics, very well. This way, someone who builds a Big Data system (in the sense described above) can create the illusion of solving the whole problem from soup to nuts even if they are focused on just a piece of it. The same goes for someone who develops a novel analysis algorithm. But Big Data is most definitely not machine learning on Map Reduce.

A group of leading researchers from across the United States wrote a whitepaper to address this misperception, see [1]. A shorter version, making the same main points, appeared in CACM, July 2014 [2]. Fig. 1 is reproduced from this whitepaper. The main point it makes is that there are many steps to the Big Data analysis pipeline, with crucial decisions required at each step, and many challenges to address in each. The first decision is what data to record or acquire, and how to make the best of data that is imperfect. Then decisions must be made to represent the data in a manner suitable for analysis, possibly after extraction, cleaning, and integration with other data sources. 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. The final interpretation step is perhaps the most crucial, because it cannot be delegated – someone is responsible for making decisions based on the result of the data analysis and this person has to understand and trust the results obtained first. Gaining this confidence will often require provenance and explanation, may need visualization, may even need sensitivity analyses of various types. All of these have to be planned for and performed effectively for the Big Data analysis to produce any real value.

4. Big Data Myth 4: Data reuse is low hanging fruit

We often have data collected for some purpose. It should be possible to use it for a different purpose as well, thereby eliminating the substantial costs of collecting data the second time. (In fact, reuse may be unavoidable in many cases, if the second analysis is performed at a later time, when there is no possibility of going back in time to collect historical data again). While this is a compelling opportunity, exploiting it requires addressing multiple challenges.

First, the original data set has to be found at the time of the desired reuse. It is relatively easy to tag data sets (or even make use of existing labels in the data set, such as attribute and table names) to find data sets that are on the topic area of interest. But,

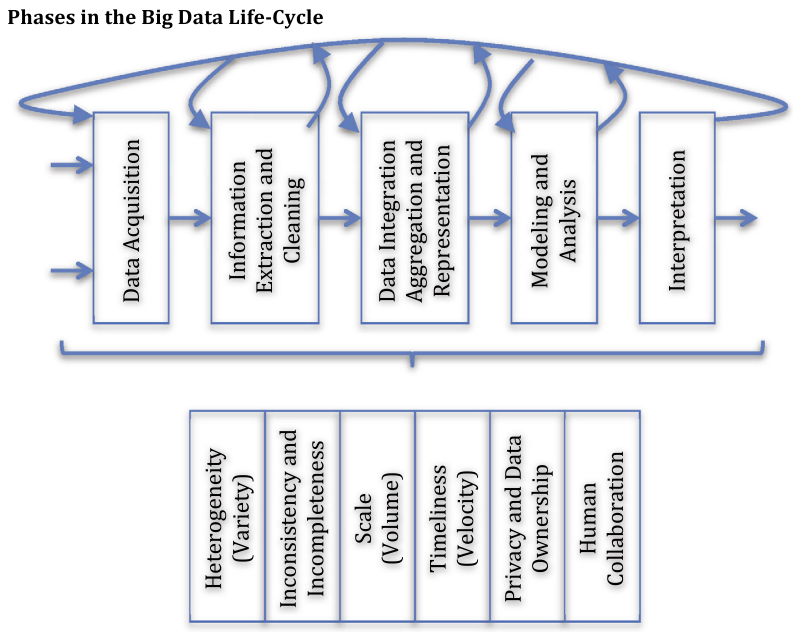

Fig. 1. The Big Data analysis pipeline. Major steps in the analysis of Big Data are shown in the top half of the figure. Note the possible feedback loops at all stages. The bottom half of the figure shows Big Data characteristics that make these steps challenging.

in a large universe of data sets, there could be hundreds of data sets that are somehow related to the topic of interest with only very few that actually have data on the relationship(s) of interest measured under conditions of interest. We are only now beginning to think about how we characterize data sets to make them findable.

Second, data sets must be understood and interpreted for them to be reusable. Obviously, this requires adequate metadata. Unfortunately, the word "adequate" in the preceding sentence is often ignored. If we know the creator and date, and the schema declaration, that is insufficient metadata in most cases. It is quite likely that it matters precisely under what conditions the data were obtained, using what instruments, after what kind of sample preparation. There is active work on metadata standards in many communities. Adhering to these standards will definitely move us forward substantially. However, we also need to address the issue of incentives, at least in the scientific community: why will a scientist spend time recording careful metadata? Why not just do the bare minimum required by the publication venue or funding agency? Furthermore, there remains sufficient diversity even within any one academic sub-discipline that many of these metadata standards do not require details that may be crucial in some specific case, even if not generally applicable. Efforts to establish a culture of data citation are crucial to address these problems.

Third, data sets found are often not quite in the right form for the desired use. Sometimes this is simply a question of performing a schema mapping. But often more substantial mismatches have to be resolved. One problem that I am currently addressing has to do with administrative data, which tend to be reported rolled up by administrative jurisdiction. When such data are reused, they need to be compared to (or joined with) data rolled up according to a different administrative hierarchy. If the two hierarchies differ, such matching is not immediately possible. For example, it is not straightforward to compare data reported by school district with data reported by county. Our approach to this problem is to develop innovative interpolation methods.

In short, data reuse is critical to address and holds out great promise. But it also poses many challenging questions, which are only now being given the required attention.

5. Big Data Myth 5: Data Science is the same as Big Data

The ability to collect and analyze massive amounts of data is revolutionizing the way scientific research is being conducted [3].

- The Sloan Digital Sky Survey [9] has transformed Astronomy from a field where taking pictures of the sky was a large part of an astronomer's job to one where the focus is on discovering interesting objects and phenomena from the databases.
- 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.
- The size and the number of experimental data sets in many applications are increasing exponentially. Consider, for example, the advent of Next Generation Sequencing (NGS) [7]. The growth rate of the output of current NGS methods is faster than the performance increase for the SPECint CPU benchmark, representing increase in computational power due to Moore's law.
- Both the volume and velocity of data require new approaches to data management and analysis. For example, the raw image datasets in NGS are so large (many TBs per lab per day) that it is impractical today to even consider storing them. Rather, these images are analyzed on the fly to produce the sequence data.

Many people use the two terms "Data Science" and "Big Data" interchangeably, applying these terms to all of the examples listed above. This is not completely inappropriate: the primary difference between the two terms is their perspective: "Big Data" begins with the data characteristics (and works up from there), whereas "Data Science" begins with data use (and works down from there). However, their formal definitions differ in more than just perspective.

The National Consortium for Data Science, an industry and academic partnership established at UNC, Chapel Hill in 2013, defines data science as "the systematic study of digital data using scientific techniques of observation, theory development, systematic analysis, hypothesis testing, and rigorous validation." A key purpose of

data science is [8] to use data to describe, explain, and predict natural and social phenomena by:

- Creating knowledge about the properties of large and dynamic data sets;
- Developing methods to share, manage, and analyze digital data; and
- Optimizing data processes for factors such as accuracy, latency, and cost.

Comparing this definition of Data Science with the Gartner definition of Big Data we saw previously, we immediately notice that it is possible to do Data Science without doing Big Data, and vice versa. Of course, nothing stops Data Science from involving Big Data, and it indeed frequently does. However, restricting our attention to the intersection of the two is needlessly limiting.

Another point to note is that Data Science tasks usually involve data analysis by a domain expert with limited database expertise. If domain expert is to succeed, data must be usable.

Unfortunately, database systems are very hard to use. There is even an urban legend about some vendors intentionally keeping them hard to use because they make so much money from consulting and support fees. In addition to the systems themselves, there are also the analysis tasks – often, we have statistically naïve users making unsupported assumptions about the data at hand, e.g. regarding independence or randomness or how representative a data set is. If we do not help people make intelligent use of their data, they will get burned and they will become opponents of all the good that our technology can bring. Database and data analytics usability research is crucial.

6. Big Data Myth 6: Big Data is all hype

Data analysis has been around for quite a while. Databases too. So what has changed? Why is now the time to get excited about Big Data? Is this merely some hype cooked up by breathless journalists?

Given the tremendous attention being paid to Big Data, this is a fair question to ask. But we see that data collection is cheap today, due to ubiquitous digitization, business process automation, the web, and sensor networks, in a way that it never was before. Data storage is cheap too, due to falling media prices. In consequence, nearly every field of endeavor is transitioning from "data poor" to "data rich." So it is not surprising that everywhere around us we have people asking about the potential of Big Data.

At the same time, we have a growing social understanding of the consequences of Big Data. We are only beginning to scratch the surface today in our characterization of data privacy. Our appreciation of the ethics of data analysis is also in its infancy. Mistakes and overreach in this regard can very quickly lead to backlash that could close many things down. But barring such mishaps, it is safe to say that Big Data may be hyped, but there is more than enough substance there for it to deserve our attention.

Acknowledgements

Supported in part by NSF Grants 1017296 and 1250880.

References

[1] Challenges and opportunities with Big Data, a community white paper available at http://cra.org/ccc/docs/init/bigdatawhitepaper.pdf.
[2] H.V. Jagadish, Johannes Gehrke, Alexandros Labrinidis, Yannis Papakonstantinou, Jignesh M. Patel, Raghu Ramakrishnan, Cyrus Shahabi, Big data and its technical challenges, Commun. ACM 57 (7) (July 2014) 86–94, http://dx.doi.org/10.1145/2611567.
[3] Advancing Discovery in Science, Engineering Computing Community Consortium, Spring 2011.
[4] Frontiers in Massive Data Analysis, National Academies Press, 2013.
[5] Pattern-Based Strategy: getting value from Big Data, Gartner Group press release, available at http://www.gartner.com/it/page.jsp?id=1731916, July 2011.
[6] The 4 V's of Big Data, http://www.ibmbigdatahub.com/tag/587.
[7] Scott D. Kahn, On the future of genomic data, Science 11 (February 2011) 728–729.
[8] Establishing a National Consortium for Data Science, available at http://data2discovery.org/dev/wp-content/uploads/2012/09/NCDS-Consortium-Roadmap_July.pdf, 2012.
[9] SDSS-III: Massive Spectroscopic Surveys of the Distant Universe, the Milky Way Galaxy, and Extra-Solar Planetary Systems, available at http://www.sdss3.org/collaboration/description.pdf, Jan. 2008.