Resources Policy 46 (2015) 234–238

Contents lists available at ScienceDirect

Resources Policy

journal homepage: www.elsevier.com/locate/resourpol

The case for “n«all”: Why the Big Data revolution will probably happen differently in the mining sector

Robert K. Perrons [a,b,*], Derek McAuley [c]

[a] Queensland University of Technology, GPO Box 2434, Brisbane QLD 4001, Australia
[b] Centre for Strategy and Performance, University of Cambridge, United Kingdom
[c] University of Nottingham, Horizon Digital Economy Research, Triumph Road, Nottingham NG7 2TU, United Kingdom

ARTICLE INFO

Article history:
Received 23 July 2015
Received in revised form
17 October 2015
Accepted 19 October 2015

Keywords:
Big Data
Predictive analytics
Mining
Information technologies
Data

ABSTRACT

Big Data and predictive analytics have received significant attention from the media and academic literature throughout the past few years, and it is likely that these emerging technologies will materially impact the mining sector. This short communication argues, however, that these technological forces will probably unfold differently in the mining industry than they have in many other sectors because of significant differences in the marginal cost of data capture and storage. To this end, we offer a brief overview of what Big Data and predictive analytics are, and explain how they are bringing about changes in a broad range of sectors. We discuss the “N=all” approach to data collection being promoted by many consultants and technology vendors in the marketplace but, by considering the economic and technical realities of data acquisition and storage, we then explain why a “n « all” data collection strategy probably makes more sense for the mining sector. Finally, towards shaping the industry’s policies with regards to technology-related investments in this area, we conclude by putting forward a conceptual model for leveraging Big Data tools and analytical techniques that is a more appropriate fit for the mining sector.

1. Introduction

“Big Data” and predictive analytics have received significant attention from the media (e.g., Harford, 2014; Lohr, 2012a, 2012b) and academic literature (e.g., Chen, Chiang, and Storey, 2012; Howe et al., 2008; Lynch, 2008; Perrons and Jensen, 2015) throughout the past few years, and it is likely that these emerging technologies will materially impact the mining sector. In fact, the industry has already made modest inroads into this area (Wilson and Hume, 2013) and, by several measures, the mining sector already has data that could fairly be characterized as “big.” Modern seismic data centers like those used by the mining and oil & gas industries can easily contain as much as 20 petabytes[1] of information, which is roughly equivalent to 926 times the size of the U.S. Library of Congress (Beckwith, 2011). If this amount of information was copied into books and put on a single continuous bookshelf, it would go around the Earth’s equator approximately six times (Beckwith, 2011). And while seismic data sets are notoriously large and cumbersome, many other aspects of the mining industry are also generating significantly more data than they used to (Annavarapu, Kemeny, and Dessureault, 2012; Bascur and Linares, 2006; Dessureault, 2007; Jahedsaravani, Marhaban and Massinaei, 2014). What is more, there is every reason to believe that this trend towards more digital information will continue to gather pace. Current estimates suggest that the total amount of digital data in the world—including things like books, images, e-mails, music, and video—is doubling every 2–3 years (Lohr, 2012a; Mayer-Schönberger and Cukier, 2013), and it seems unlikely that the mining sector will not be affected by this.

This short communication argues, however, that these technological forces will probably unfold differently in the mining industry than they have in many other sectors. To this end, we begin by offering a brief overview of what Big Data and predictive analytics are, and explain how they are bringing about changes in a broad range of sectors. Then we will discuss the “N=all” approach to data collection and information management—that is, eschewing the sampling of data in favor of using all of the data pertaining to a particular phenomenon—which is frequently advocated by many of the consultants and technology vendors offering solutions in the Big Data domain (Harford, 2014). Next, based on high-level discussions with other industries and early observations about Big Data’s nascent advances into the mining sector, we will make the case for

* Corresponding author at: Queensland University of Technology, GPO Box 2434, Brisbane QLD 4001, Australia.
E-mail address: perrons@alum.mit.edu (R.K. Perrons).

[1] This statistic is less impressive when you consider that Walmart, the U.S. retail giant, collects more than 2.5 petabytes of data every hour from customer transactions (McAfee and Brynjolfsson, 2012).

http://dx.doi.org/10.1016/j.resourpol.2015.10.007
0301-4207/

adopting a "n « all"[2] data collection strategy within the mining industry instead of the "N=all" approach. Finally, towards shaping the sector's policies with regards to technology-related investments in this area, we conclude by putting forward a conceptual model for leveraging Big Data tools and analytical techniques that is a more appropriate fit for the mining industry.

2. What is Big Data? And how can it deliver value?

The term "Big Data" is a rather vague label that describes the application of new tools and techniques to digital information on a size and scale well beyond what was possible with traditional approaches (Lohr, 2012b). The idea has grown organically from a broad range of academic disciplines, technology providers, and consultants and, as a consequence, no single definition has been agreed for this steadily evolving research area. Despite the absence of a precise definition, however, most users of the term broadly agree that it refers to data sets and analytical techniques in applications that are so large and complex that they require advanced data storage, management, analysis, and visualization technologies (Chen et al., 2012).

At its core, the main objective of Big Data and predictive analytics is something that organizations have been aspiring to for a long time: to make better decisions (Regalado, 2014). This is achieved by statistically processing large amounts of data—either structured or unstructured—from disparate sources that are often left untapped by conventional business intelligence programs in order to uncover hidden patterns, unknown correlations, and other types of useful information. These can potentially result in probability-based actionable insights that help human decision-makers to figure out what to do next. These new capabilities have most famously delivered value in the social media sector and retail industries (Harford, 2014; Mayer-Schönberger and Cukier, 2013), but have also led to major breakthroughs in a diverse range of other contexts, including scientific research (Frankel and Reid, 2008; Kluger, 2014; Lynch, 2008), healthcare (Chen et al., 2012), and professional sports (Leahey, 2013). And Big Data has also made significant forays into the heavy equipment sector—an industry that directly supplies the mining community. As Mehta (2013) notes, Caterpillar now uses Big Data in the short-term to anticipate repairs and more efficiently plan component inventories and, in the long-term, to improve product designs and predict changes in their customers' needs.

By stark contrast, the mining industry has made relatively modest progress in the Big Data domain (e.g., Wilson and Hume, 2013). But these analytical approaches could almost certainly deliver considerable value in that industry. For example, large haul trucks have been a mainstay of the mining industry for a long time (Santos et al., 2010), and a significant fraction of the sector's workplace fatalities occur in and around these kinds of large vehicles (Groves et al., 2007). Large data sets containing geospatial information about the pathways of haul trucks could be compared to synchronized accelerometer data measured from inside each truck to tell road maintenance crews where the most egregious potholes and depressions are appearing (Mednis et al., 2011), thereby making it possible to identify potentially hazardous parts of the road before they become more serious.

[2] Whereas a single "<" symbol denotes in mathematics that something is less than something else, the "«" symbol means that it is much less. Also, the distinction between the upper-case "N" in "N=all" and the lower-case "n" in "n « all" throughout this paper is intentional. N typically denotes the size of the entire population, while n represents the size of a sample that is smaller than N within that same population (Levin, 1987).

The Big Data and predictive analytics trend can be seen as the convergence of a number of new ways to extract value from business data. Specifically, these include:

- The continued decrease in the cost of technology for capturing, storing, and processing data leading to nearly all business processes becoming digital and generating rich business data
- The move to make these diverse datasets available in real-time within the business for new and creative uses, together with the increasingly available option of augmenting these data sets with open and public data from other sources
- The use of computationally intensive analytic techniques to obtain new quantitative insights from previously unexplored combinations of data that, at first blush, may seem to have no obvious relationship to each other (Mayer-Schönberger and Cukier, 2013). In some circumstances, simply finding such correlations is sufficient to release the value. For many other scenarios, however, these kinds of statistical insights need interpretation by someone who understands the mechanisms behind the phenomena being observed, as correlation is not the same as causality. This is a major issue, for example, in Big Data-based medical research. A team of researchers from Google announced in the high-profile journal *Nature* that, by closely tracking search terms like "flu symptoms" or "pharmacies near me," they were able to predict the spread of influenza through the U.S. more quickly and accurately than the Centers for Disease Control and Prevention (Harford, 2014). After a few years of providing reliable estimates, however, the model's predictions became much less reliable because "Google did not know—could not begin to know—what linked the search terms with the spread of flu" (Harford, 2014).

To date, many approaches to Big Data and predictive analytics have involved bringing these kinds of digital information together within a single data centre, building so called "data lakes," and then processing the data using cluster-computing software such as Hadoop (Shvachko et al., 2010). For the mining sector, such an environment represents a particularly good vantage point from which to investigate the widest range of possible data interactions, especially in light of the fact that the off-the-shelf software implementing the various analytic techniques has been designed to run efficiently in this type of scenario. Also, with the relevant data in one place, the software is not constrained by arbitrary data partitioning—that is, the process of determining which data subjects, data occurrence groups, and data characteristics are needed at each data site in a geographically distributed system (Wu et al., 2013).

3. "N=all" and its limitations

One of the major tenets of Big Data solutions being prescribed by many consultants and technology vendors is the concept of "N=all"—that is, to collect and analyze all of the data potentially pertaining to a particular phenomenon. The notion of taking a sampled subset of representative data to identify and measure trends within larger populations has been a cornerstone of many analytical techniques for generations (Cooper and Emory, 1995; Levin, 1987) but, as Mayer-Schönberger and Cukier (2013) suggest:

> Sampling is an outgrowth of an era of information-processing constraints, when people were measuring the world but lacked the tools to analyze what they collected… However, the concept of sampling no longer makes as much sense when we can harness large amounts of data. The technical tools for handling data have already changed dramatically, but our methods and