Strategic Directions for Big Data Analytics in E-Commerce with Machine Learning and Tactical Synopses: Propositions for Intelligence Based Strategic Information Modeling (SIM)

Jim Samuel
William Paterson University

Rajiv Kashyap
William Paterson University

Stephen Betts
William Paterson University

E-commerce has seen tremendous growth in big data and the continued acceleration in growth of information facets. There has been a significant scaling up of information quantity, granularity, complexity, equivocality and variety. Data analytics tools and techniques (DATT) such as machine learning and artificial intelligence have been widely leveraged to gain competitive advantage and such resources are readily available. However, there has been a lack of clarity surrounding, what we term as 'strategic information modeling' (SIM). Our research presents propositions to provide contextual clarity to the rapidly expanding Big Data environment and also an articulation of the emerging informational challenges in e-commerce. Our analysis provides insights into the potential role of SIM and SIM generated competitive advantages and concludes with e-commerce relevant propositions for an optimal path towards SIM and machine learning.

INTRODUCTION

> "Not everything that can be counted counts, and not everything that counts can be counted."
> - Albert Einstein, Physicist.

In the age of 'Big Data', it has become stylish to claim to have more data than can be processed and an ability to project petabyte growth rates dwarfing existing data processing infrastructure to be a sign of superior sophistication. 'Big Data' has been one of the most searched for phrases across search engines for the past five years, however the notion of Big Data has expanded beyond the obvious association with the quantity, and Big Data definitions have been evolving with technology, research and practice. Early articulation defined Big Data with the three V's of data 'Volume', data type 'Variety' and data speed 'Velocity' (Russom, 2011), which built upon Gartner's four V's with data 'Veracity' being the fourth, though Gartner (Laney, 2001) framed the four V's in a limited context. A Forbes writer titled an amusing article "12 big data definitions –what's yours?" (Press, 2014), highlighting the range of available and emerging Big Data definitions. Recent research has identified distinguishing properties of Big Data based on "exhaustivity, resolution, indexicality, relationality, extensionality and scalability" (Kitchin & McArdle, 2015), much of which serve as indicators of the complexity surrounding Big Data phenomena. Essentially, along with intrinsic data characteristics, the role of context, temporality, objectives, technological change, relativism (data dimensions contrasted to data processing capabilities), belief and behavior are all factored in to craft weighted perspectives and definitions of Big Data. Each term that has come to be and created to be associated with Big Data reflects the complex challenges facing Big Data based analytics and decision making. A fifth 'V' used to describe BIG Data is 'Value' (Wixom et al., 2013), which addresses the economic value creation potential of Big Data.

Practitioners and researchers have been working on addressing Big Data challenges using a host of data management, technology architecture, distributed computing, and automation techniques. 'Big Data Analytics' (BDA) techniques are prominently used to explore massive quantities of data to gain actionable insights. BDA has been defined as: being the place "where advanced analytic techniques operate on big data" (Russom, 2011) and as being "a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling high-velocity capture, discovery, and/or analysis.' (Gantz and Reinsel, 2011). BDA has also been described by defining sub-domains such as structured data analytics, text analytics, web analytics, multimedia analytics, network analytics, mobile analytics (Hu et al., 2014), semi-structured and unstructured data, applied domain based (business, environment, social, etc;) and objective driven. BDA contextualized to our present research perspectives is defined as being an evolving collection of a broad range of processes, techniques, technologies, domain expertise and quantitative models used to process massive data, static and streaming, with varying facets, dimensions, granularity, and characteristics.

The present research, from an ontological perspective, also emphasizes the temporality and relativity attributes of BDA – much of BDA is driven by the changing characteristics (temporality) of data and associated technologies, and the lag between the complex challenges presented by massive data and technology driven BDA solutions being developed. As businesses increasingly consume BDA driven information, it is imperative for them to be effective in ensuring high quality and timeliness of their BDA deliverables. As these are highly competitive in nature, each entity strives to reduce the lag and do so effectively. Our research aims to provide a strategic model for 1) prioritizing data with high potential for value creation, 2) reducing lags between evolving challenges and nascent solutions, and 3) improving the economic value of BDA. The sections that follow discuss 'internet of things' (IoT), BDA challenges in e-commerce, machine learning, benefits and challenges of existing DATT, strategic information modeling (SIM) and the scope for SIM driven competitive advantages and concludes with propositions for optimal SIM using machine learning concepts.

LITERATURE REVIEW

One of the dominant contributors to the Big Data phenomenon is the most recent massive technology wave, which is already permeating human society at multiple levels via the 'internet of things' (IoT). International Data Corporation has estimated that strong IoT growth at a compounded annual growth rate (CAGR) of 20.0% in corporate expenditures will reach $7 trillion in 2019 (Lund, et al., 2014). Other studies have indicated that the world will have over 100 billion interconnected IoT devices with a financial influence of above $11 trillion by the year 2025 (Rose, et al., 2015). IoT has been defined as "a type of computing characterized by small, often dumb, usually unseen computers attached to objects. These devices sense and transmit data about the environment or offer new means of controlling it" (MIT TR, 2014). Such technological developments will generate fresh avenues for value creation in most sectors including e-commerce, finance, education, social projects, travel, automobiles, "e-health, retail, green energy, manufacturing, smart cities/houses and also personalized end-user applications" (Barnaghi et al., 2012). IoT technology development trends indicate a move towards highly interconnected innovative services, applications, products and platforms which are algorithmically driven to collect, exchange, store, manage, analyze, learn from and propagate raw and analyzed data from the plethora of

IoT devices, sensors and readers ubiquitously permeating the physical environment globally. With such immense data generation and hyper-connectivity potential, IoT in e-commerce will add scope and complexity to the present BDA challenges and new strategies will be necessary to address the same.

There are multiple fronts to BDA challenges facing ecommerce companies – the need to use BDA in creating operational efficiencies, evolving customer expectations and associated BDA driven effectiveness, BDA driven competitiveness, security issues associated with IoT and BDA, continued adoption of technological advancements in IoT and BDA, and the innovation potential of BDA. Ecommerce companies that have employed BDA have seen a relative benefit of 5%-6% in their productivity (McAfee et al., 2012). In its recent report McKinsey (2016) identified ecommerce as one of the key domains that stands to achieve significant benefits from BDA - global e-commerce giants Alibaba and others have used BDA to manage their microloans offerings by relying on real-time data processing and analytics. Early studies have shown multiple avenues for BDA drive value creation in ecommerce through population segmentation, personalization, operating margins, value-add offers, product development, strategic insights and innovation (Miller, 2012). However, statistics and surveys also point to a significant amount of Big Data that ecommerce companies are unable to leverage due to growing data volumes and associated complexities, lack of resources and limited computing power, indicating significant 'uncaptured value' (McKinsey, 2016). Many of the generally known business challenges are accentuated in the context of e-commerce where non-digital social interaction is extremely limited and often missing altogether. Often, decision making within e-commerce is devoid of the physical interaction dynamics and is largely dependent upon an organization's ability to make high quality decisions based on BDA. It is impossible to be competitive in such a rapidly expanding information environment without advanced domain sensitive data management tools and strategic information modeling. Thus, huge potential for value creation remains untapped and though BDA is continuously improving with respect to mathematical modeling, programming and computing resources, existing solutions continue to trail the opportunities presented by evolving Big Data.

Machine Learning (ML) consists of "computer systems that automatically improve with experience" which caters to "the demand for self-customizing software" and has significant application in speech recognition, computer vision and graphics, robotics and wide variety of empirical, deterministic and stochastic models based applications (Mitchell, 2006). Industry experts have called for business leaders to quickly adopt ML techniques and ML based strategies to achieve competitive advantages. They have also warned that companies that fail to leverage ML, optimization and artificial intelligence technologies will fail to be competitive and soon become "legacy companies" (Pyle and Jose, 2015). ML methodologies have been used to improve ecommerce competitiveness significantly in multiple areas of search, personalization and customization, shipping and transportation, logistics optimization, fraud detection, optimization, cost reduction, sentiment analytics, recommendation mechanisms, pricing and cross-selling (Michael and Mitchell, 2015). Furthermore, ML has the potential to create additional value as ML feeds on learning from experience and large troves of data remain unleveraged and underutilized due to resource and skill constraints (McKinsey, 2015). The need for ML driven solutions combined with the rapidly evolving data dynamics require looking beyond granular mathematical modeling and software programming for higher lever constructs that could constitute strategic solutions.

PROPOSITION DEVELOPMENT

Our objective here is to identify conceptual constructs which can be used to create a layer of intelligence which will serve as a "Strategic Information Modeling" layer for relatively efficient implementation of BDA without significant increase in computing resources. Based on our review of domain literature and practitioner focus, we identify five primary variables into which we can conceptually converge most the existing BDA dynamics: Data, Computing Power, Analytics, Quality and Actionability. Data (" d ") refers to the massive, structured and unstructured, voluminous data pouring in at high speeds, with varying characteristics and qualities, and to the systems that are used to store such data. Computing power (" p ") refers to the cumulative data processing capabilities, distributed-hybrid-centralized, the various layers of computational infrastructure for acquiring, managing the storage of and the application of techniques upon the stored data. Analytics (" i ") refers to the collective portfolios of mathematical models, software script, statistical procedure, data-specific processes, domain specific lifecycle practices, objectives and human skills. Quality (" q ") refers to the classical information relevant constructs of timeliness, relevance, completeness, accuracy and certainty, as applicable to the output of BDA. Actionability (" r ") refers to the properties of the BDA driven information which is clear, quantified, objective and ready for use as reliable input for successful decision making. The challenge of present day BDA is that for BDA value (" V ") to be captured or generated, the vector of which is indicated by:

$$V \rightarrow f(q, r) \tag{1}$$

It is necessary to invest vast amounts of effort (" E "), the vector of which is indicated by:

$$E \rightarrow f(d, p, i) \tag{2}$$

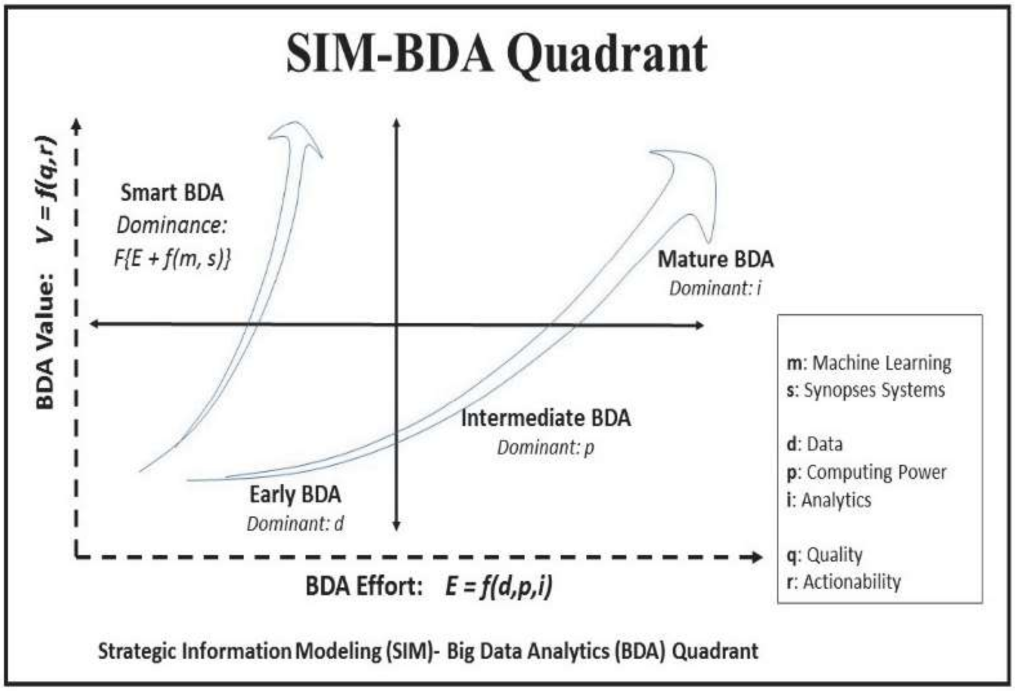
Where V is positively dependent on E

$$V \alpha E \tag{3}$$

This leads to a gradual yield of value, while significant effort is invested in BDA. In what we term as the Evolutionary BDA Model, companies have evolved through Early, Intermediate, and Mature BDA stages – the latter yielding the most potential for value creation and competitive advantages.

Early stage BDA is dominated by the sheer volumes, velocity, variety and veracity properties of data. For instance, large e-commerce players like Dell and Amazon initially made massive investments in IT infrastructure (data collection, storage, human resources). Access to data was the primary driver of competitive advantage. Intermediate BDA is dominated by the acquisition of computing and processing capabilities. Propelled by the scale of their IT infrastructure, these e-commerce companies were able to leverage massive computing power to integrate platforms, business processes, and markets to obtain pre-emptive and first mover advantages. Mature BDA is dominated by 'deep analytics' models and mathematics. At this stage companies leverage their expertise due to prior experience and learning from BD analysis. During mature BDA, companies draw upon their deep learning to create innovative solutions for their partners. They cement relationships by monetizing their analytics driven capabilities that allow partners to create value and generate new revenue streams (e.g. GE's Predix and Predictivity systems – see Iansiti and Lakhani 2014). Furthermore, this approach has limited efficiencies since the changing characteristics of data and decision quality can impact companies and organizations as they move from early BDA to intermediate BDA, and finally to Mature BDA where significant value creation potential exists.

FIGURE 1
STRATEGIC INFORMATION MODELING DRIVEN COMPETITIVE ADVANTAGES



In contrast, we posit that smart BDA fueled by ML and synopses systems has the potential to leapfrog the Evolutionary BDA Model. Our research propositions a macro-level shift in approaching BDA. We posit that, with a relatively lower investment of effort, it would be possible to surpass the value capture potential of BDA using our northwest quadrant “Smart BDA” (Figure 1) strategy. In this quadrant, we introduce two constructs for accentuated value capture or generation: Machine learning (as “ m ” in equations, ML in abbreviated text) as a portfolio of techniques, and tactical synopses systems (“ s ”) which refers to data management techniques and optimization tools which allow the essence of massive data to be captured in a reduced but higher degree relevance structure using artificial intelligence tools focused on critical knowledge extraction (Mani et al., 2014 and Cormode et al., 2012). This approach of combing ML with tactical synopses eliminates the classical arguments against using pure “synopses” of data as such simplistic approaches could ‘lose the “long tail” in dataset’ (Cohen et al., 2009).

PROPOSITIONS

Present approaches to BDA involve a sophisticated mix of early stage emphasis of collecting massive amounts of data and tackling the efficient storage of the same, with intermediate and mature BDA stages which iterate between emphasizing computing infrastructure and analytics solutions. All of this comes at a tremendous strain on the economics, human resources and competitiveness of organizations –our analysis of the key drivers involved in each stage of BDA leads us to the proposition that, ceteris paribus (everything else being equal) in a BDA environment.

Proposition 1

Additional Value (“ Vs ”) can be created by using ML (“ m ”) in conjunction with tactical synopses (“ s ”) with relatively lower investments of data, computing power and advanced analytics resources.

This implications of this proposition are expressed in the equations below:

$$Vs \rightarrow f(m, s) \tag{4}$$

and for total value or effective value (“ Vt ”) at relative time n:

$$Vt(n) \rightarrow F(E(n) + f(m, s)) \tag{5}$$

Given the relatively lower focus on extensive collection of massive data and computing power, along with streamlined analytics, we posit that such Vs can be generated with greater time efficiency and this lead to better timeliness of actionable insights:

Proposition 2

Additional Value (“ Vs ”) can lead to actionable BDA with greater timeliness than intermediate or mature stage BDA.

The above propositions accommodate can accommodate a wide range of structured and unstructured data and can thus be used universally for BDA. The propositions also account for changes in computing power, relativeness of time and depth and breadth of analytics tools, models and techniques. The four quadrants of the SIM-BDA quadrant model represent the various potential stages of BDA and we argue for the superiority of the north-west quadrant which is termed as the “Smart BDA” quadrant. The Smart BDA quadrant represents a the space of smart choices in BDA where the goal is to optimally maximize the cost-benefit trade-off in BDA.

DISCUSSION

There are various potential methods for empirical evaluation of the Smart BDA propositions, especially in E-Commerce. In its present form, our research is being presented as an essay based on a macro level analysis and theoretical integration of key concepts based on literature review and our own experience in BDA. Our goal has been to identify effective ways of tackling the continuously expanding BDA challenges without compromising on the quality of the potential insights and value that could captured or created through present day high-efforts based comprehensive BDA strategies. In doing so, our reference point for practical applications has been E-commerce (Miller, 2012). BDA has been widely employed in E-commerce as discussed above and we suggest the following avenues for hypothesis generation and empirical testing – big data used for segmentation, personalization, value-add offers, and recommendation can be used to evaluate both the intrinsic efficacy of the propositions as well as create a contrast between SIM-BDA based output and classical BDA output on quality and timeliness. While the above recommendations for empirical implementation are largely focused on the B2C segment of E-Commerce, it must be noted that the more voluminous B2B segment which has the biggest share of annual global E-Commerce revenues, also presents rich opportunities for empirical analysis (Fensel et al., 2001). The presence of organized players with larger scope for systematic orders in B2B, and the cyclicality, trends and statistical associations than can be established within B2B E-Commerce suggest greater potential for successful application of the SIM-BDA strategies – the potential and the need for capturing latent value in big data is a well-established need across domains (McKinsey, 2016).

LIMITATIONS

The limitation of the present study is the absence of empirical development, but this is not unusual for direction setting research that initializes fresh discussions by positing conceptual arguments. The study has drawn insights from multiple domains (information systems, data science, e-commerce, computer science, analytics and management) and may therefore lack linguistic adherence to the terms specialized to any single domain. The study has also not discussed domain specific opportunities and challenges.

Thus, the relevance to domains such as social media, healthcare and others would have been useful, but would make the articulation arduously voluminous.

In spite of these limitations, practitioners stand to gain fresh perspectives as the propositions constitute out-of-the-box thinking and provide directionality for achieving quicker results than through classical BDA strategies. Additionally, the rise of IoT and its popularity couples with its potential to multiply presently know data velocities and volumes present critical BDA challenges which cannot be efficiently addressed using classical BDA (Riggins and Wamba, 2015). The SIM-BDA four quadrant model is a very relevant model for practitioners to apply, especially given limited data acquisition and computing resources.

CONCLUSION

Our propositions present rich opportunities for future research – there is already a great emphasis on the classical dimension of BDA which involves the study of data types, big data acquisition, efficient storage and management of massive data, centralized and distributed processing mechanisms, scalability, analytics models, mathematics, statistical procedures, ML and artificial intelligence technologies (Hu et al., 2014). However as identified by McKinsey (2016), there is a significant gap between what existing methods are able to leverage and the value that remains to be captured. It is this gap that our research seeks to reduce and our propositions open a novel direction for research based on ML and tactical synopses – research opportunities based on the SIM-BDA propositions can be crafted from objective selections of a mix of variables used in the present study.

Smart BDA quadrant implementation provides a strategic alternative for organizations who are concerned about the significant investment of effort into BDA, which for many organizations is an explorative exercise based on peer pressure or a stylistic endeavor driven by brand and industry forces. Smart BDA is an ideal solution as it does not detract from the classical BDA trajectory but provides a side-stop based on additional but limited extension of resources with the early BDA stage. It therefore represents a relatively low risk endeavor with the potential for gaining high quality actionable insights in a timely manner at a relatively low effort stage. Smart BDA is not to be viewed as a comprehensive replacement for BDA but as a strategic accompaniment to existing BDA to streamline effort and improve the speed and quality of decision making.

DISCLOSURE

This paper is an expanded version of Samuel, J., Kashyap, R., & Betts, S., (2017): “Machine Learning And Tactical Synopses For Big Data Analytics In E-Commerce: Propositions For Strategic Information Modeling (SIM)”, 44th NBEA Annual conference proceedings, 2017 p227-231, with overlapping content.

REFERENCES

Akter, S., & Wamba, S. F. (2016). Big data analytics in E-commerce: a systematic review and agenda for future research. Electronic Markets, 26(2), 173-194.

Barnaghi, P., Wang, W., Henson, C., & Taylor, K. (2012). Semantics for the Internet of Things: early progress and back to the future. International Journal on Semantic Web and Information Systems (IJSWIS), 8(1), 1-21.

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. MIS quarterly, 36(4), 1165-1188.

Cohen, J., Dolan, B., Dunlap, M., Hellerstein, J. M., & Welton, C. (2009). MAD skills: new analysis practices for big data. Proceedings of the VLDB Endowment, 2(2), 1481-1492.

Columbus, L. (2014). Making analytics accountable: 56 % Of executives expect analytics to contribute To 10 % Or more growth in 2014. Forbes.

http://www.forbes.com/sites/louiscolumbus/2014/12/10/making-analytics-accountable-56-of-executives-expect-analytics-to-contribute-to-10-or-more-growth-in-2014/

Cormode, G. (2013, July). Summary data structures for massive data. In Conference on Computability in Europe (pp. 78-86). Springer Berlin Heidelberg.

Cormode, G., Garofalakis, M., Haas, P. J., & Jermaine, C. (2012). Synopses for massive data: Samples, histograms, wavelets, sketches. Foundations and Trends in Databases, 4(1–3), 1-294.

Fensel, D., Ding, Y., Omelayenko, B., Schulten, E., Botquin, G., Brown, M., & Flett, A. (2001). Product data integration in B2B e-commerce. IEEE Intelligent Systems, 16(4), 54-59.

Gantz, J., & Reinsel, D. (2011). Extracting value from chaos. IDC iview, 1142(2011), 1-12.

Hu, H., Wen, Y., Chua, T. S., & Li, X. (2014). Toward scalable systems for big data analytics: A technology tutorial. IEEE Access, 2, 652-687.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255-260.

Kitchin, R., & McArdle, G. (2015). The diverse nature of big data.

Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. META Group Research Note, 6, 70.

Lund, D., MacGillivray, C., Turner, V., & Morales, M. (2014). Worldwide and regional internet of things (iot) 2014–2020 forecast: A virtuous circle of proven value and demand. International Data Corporation (IDC), Tech. Rep.

Mani, G., Bari, N., Liao, D., & Berkovich, S. (2014, August). Organization of knowledge extraction from big data systems. In Computing for Geospatial Research and Application (COM. Geo), 2014 Fifth International Conference on (pp. 63-69). IEEE.

McAfee, A., Brynjolfsson, E., Davenport, T.H., Patil, D.J. and Barton, D., (2012). Big data. The management revolution. Harvard Business Review, 90(10), 61–67.

McKinsey report: "The Age Of Analytics: Competing In A Data-Driven World", December 2016

Miller, G., (2013). 6 ways To use "big data" To increase operating margins By 60 %. Available at: http://upstreamcommerce.com/blog/2012/04/11/6-ways-big-data-increase-operating-margins-60-part-2

MIT Technology Review, "Internet of Things" – P1, Julu/August 2014.

Mitchell, T. M. (2006). The discipline of machine learning (Vol. 9). Carnegie Mellon University, School of Computer Science, Machine Learning Department.

Pelino, M and Gillett, F.,"The Internet Of Things Heat Map, 2016: Where IoT Will Have The Biggest Impact On Digital Business", Forrester Report, 2016

Press, G., 2014, https://www.forbes.com/sites/gilpress/2014/09/03/12-big-data-definitions-whats-yours/print/

Pyle, D. and Jose, C.S., "An executive's guide to machine learning" http://www.mckinsey.com/industries/high-tech/our-insights/an-executives-guide-to-machine-learning

Riggins, F. J., & Wamba, S. F. (2015, January). Research directions on the adoption, usage, and impact of the internet of things through the use of big data analytics. In System Sciences (HICSS), 2015 48th Hawaii International Conference on (pp. 1531-1540). IEEE.

Rose, K., Eldridge, S., & Chapin, L. (2015). The internet of things: An overview. The Internet Society (ISOC), 1-50.

Russom, P. (2011). Big data analytics. TDWI best practices report, fourth quarter, 19, 40.

Wamba, S. F., Akter, S., Edwards, A., Chopin, G., & Gnanzou, D. (2015). How 'big data'can make big impact: Findings from a systematic review and a longitudinal case study. International Journal of Production Economics, 165, 234-246.

Wixom, B. H., Yen, B., & Relich, M. (2013). Maximizing Value from Business Analytics. MIS Quarterly Executive, 12(2).