Cite: Abubakar, A.M. (2019). Big Data Analytics: A State-of-the-Art Review. 2nd International Conference on Data Science and Applications (ICONDATA'19) Edremit, Balıkesir / Turkey

Big Data Analytics: A State-of-the-Art Review

Dr. Öğr. Üyesi. A.Mohammed ABUBAKAR

Antalya Bilim Üniversitesi, İktisadi İdari ve Sosyal Bilimler Fakültesi, Antalya, Türkiye

Mohammed.abubakar@antalya.edu.tr

Özet

Büyük veri analizi, müzakere, fikir çatışma ve tartışmalara konu olmuştur. Bununla birlikte, büyük verilerin üç farklı bakış açısından (yani veri tanılaması, veri çeşitliliği ve veri yönetişimi) uygulanabilirliği ve dezavantajlarına rağmen, aralarındaki ilişkiyi inceleyen çalışmalar ilginç bir şekilde sınırlı düzeydedir. Bu çalışmada, bilgi değeri doğrultusunda veri tanılaması, veri çeşitliliği ve veri yönetişimi hakkında kısa bir genel değerlendirme sunulmaktadır. Bu konu esasıyla, bu çalışmada büyük veri kullanımının karşı karşıya kaldığı enteresan ve önemli konuları gündeme getirmekte ve acil dikkat gerektiren bir dizi araştırma sorusu ile sonuçlanmaktadır.

Anahtar Kelimeler: *Büyük veri, veri tanılaması, veri çeşitliliği, veri yönetişimi*

Big Data Analytics: A State-of-the-Art Review

Abstract

Big data analytics has been a subject for debate, discussions and arguments. However, the applicability and challenges of big data in terms of three views (i.e., data diagnosticity, data diversity and data governance) has been widely ignored. This paper provides a brief overview for data diagnosticity, data diversity and data governance in line with information value. In essence, this paper raises interesting and importance issues facing big data usage and concludes with a number of research questions that needs urgent attention.

Keywords: *Big Data, data diagnosticity, data diversity, data governance*

1 Overview

The popularity of the Internet and the progressive transformation of World Wide Web from Web 1.0 to Web 5.0 technologies has created a shift in Web contents creation, from publisher-to user-created contents. More and more people are using the Internet and social media outlets such as Instagram, Twitter, Facebook, Pinterest etc. to generate massive amount of data within a short period of time. Huge amount of data is generated by more than a billion people on daily basis is referred to as "Big Data". According to Ghasemaghaei et al. (2017), Big data analytics delineate the use of new generation of analytical tools (e.g., Python, R) empowered to analyze data that is high in terms of Volume, Velocity, Variety, Veracity and Value, popularly known as the 5V's (Volume, Velocity, Variety, Veracity and Value) as shown in figure 1

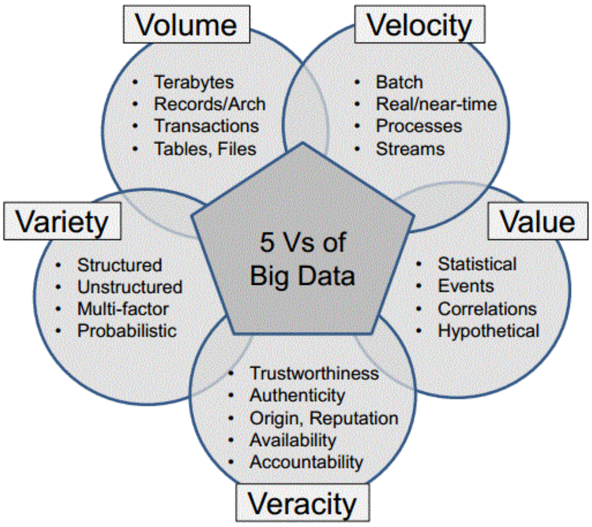

Figure 1: Big Data Five V's (Source: Sumanth, 2019)

Big data analytics has been deployed in several areas of business such as security and surveillance, marketing, human resource management, health care, e-government, education, e-commerce etc. This is because Big data analytics allow firms to discover invisible associations, hidden patterns, unknown market and business intelligence in large datasets (Chen, Preston, & Swink, 2015). Insights generated via data-driven approach can be categorized into three namely: descriptive, predictive and prescriptive. Descriptive insight stresses the importance of comprehending past activities and events e.g., firms deploy historical data to identify patterns of sales, production, customer preferences etc. This can be achieved via dashboards, scorecards and visual analyses (Sivarajah et al., 2017).

Predictive insight stresses the importance of comprehending the patterns of future events (Ghasemaghaei et al., 2016). Firms can analyze the associations between data to forecast future probabilities and trends e.g., statistical and forecasting models help firms to forecast future sales, price change and/or change in consumer preferences (Deka, 2016). Prescriptive insight stresses the importance of finding the best course of actions to get the optimal results for taking advantage of a circumstance (Appelbaum, Kogan, Vasarhelyi, & Yan, 2017).Firms can use simulative models to determine alternative business scenarios and solutions e.g., introduction of new products to specific customer segments because preference for that product has increased or using simulations to find optimal ways to reduce cost and deliver quality services (Ghasemaghaei & Calic, 2019b).

In sum, these capabilities allow firms to improve products and/or services design, enhance service delivery and operations, data driven decision-making in terms of innovation and design. Corporations and enterprises cultivate, and process data characterized with the 5Vs mentioned above, since data driven-decisions rely on data quality, poor data is equiprobable to poor decision-making (Marsden & Pingry, 2018) and wrong decisions resulting from poor data quality can cause loss in revenue that could exceed billions of dollars per year (Hazen et al., 2014). In fact, the value of data is determined mainly by its quality (Hazen et al. 2014). According to Kwon, Lee and Shin (2014) and Wang and Strong (1996), data quality is the extent to which data matches it use as required by the users. To overcome these challenges, three dominant views on how to generate informational value from data were developed (Grange, Benbasat & Burton-Jones, 2019). The first view stresses the importance of data diagnosticity (Cai & Zhu, 2015); the second emphasis is on data diversity (Grange et al., 2019) and the last but not the least, is on data governance (Mikalef & Krogstie, 2018). This study contributes to the literature by proposing triad model which encompass the above said views. See figure 2

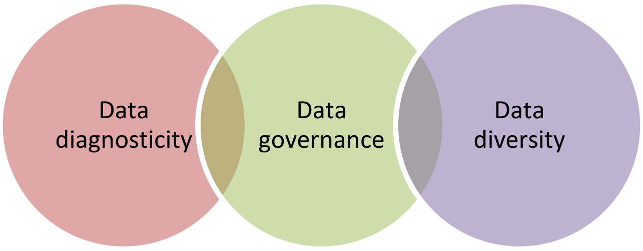

Figure 2: Big data 3'views

2 Data diagnosticity

Diagnostic data delivers highly valid, reliable, and interpretable information, which is valuable for all stakeholders e.g., organizations, top management, partners, customers and the society. The ability of business enterprises to obtain valid, reliable, valuable information and insights from large datasets is known as data diagnosticity (Grange et al., 2019). It also refers to the systematic retrieval of insights from dataset to make accurate and reliable business interpretations. Successful business enterprises are reliant on Big data analytics for insights prior to decision-making, because they analyze data to understand past and present events, and then act and use the data for future prediction.

In their influential work, Ghasemaghaei and Calic (2019a) described how credit-card firms leverage on big data to determine potential customers, "use ready market to make personalized offers in milliseconds, or they can optimize offers over time by tracking responses to predict future decisions" (p.40). Scholars and practitioners asserted that firms are skeptical about big data as they are mainly concerned about the quality of data cultivated could have adverse effects on data diagnosticity (Cai & Zhu, 2015; Kwon et al., 2014). This highlight the importance of data diagnosticity relative to informational value, what remains unclear is how to diagnose and retrieve insight from heterogenous, semi-structured, unstructured, biased, messy, and uncomplete datasets.

3 Data diversity

Diverse data delivers data that is heterogeneous subsuming varying level of views, opinions and recommendations. Bakshy et al. (2015) claimed that diverse data ensures that firms and corporations are not exposed to "filter bubbles" or "echo chambers,". In other words, the diversity of data ensures that one is not subjected to similar viewpoints, but to heterogenous views that could enhance holistic analyses. In this view, curating and having an in-depth understanding of a phenomenon on broader basis supports effective decision-making. Contrariwise to this positive view, Cai and Zhu (2015) argued that the challenge in dealing with diverse data is the level of complication especially in formatting and processing such data. More to this, diverse data sources amplify and makes it difficult to ascertain data accuracy because data comes from different sources, in different sizes and formats. According to Ghasemaghaei, Ebrahimi and Hassanein (2018), advanced data analytics tools have the capacity to spot and diagnose fruitful information from voluminous and heterogeneous data, the eventual outcome, value of managing and handling, analyzing and interpreting big data will be influenced by its quality. What remains unclear is how to diagnose and retrieve insight from diverse, unformatted, varying sizes datasets.

4 Data governance

Although big data has vast titanic benefits, it also has potholes that threaten it usage. Which is the third view that stresses the importance of data governance. Big data comprises of huge datasets – numeric and non-numeric observations, insights and intelligence curated from or about, human subjects and their surroundings. Implications for confidentiality, data abuse and misuse may arise because personal data is involved which is associated with privacy (Olhede & Wolfeb,2018). In this view, big data governance seems to play an important role in how data is collected, shared, and used by corporations.

According to Weber et al. (2009), Big data governance comprises of different activities involving decision-making for data quality management; and the allocation of responsibilities (structural practices), the decision-making process (procedural practices), as the relational responsibilities and links between departments, stakeholders, customers and products (relational practices). It therefore summarizes as the capacity of firms to obtain, store, process, analyze and interpret big data richness and combine it with real-time data to make forecast future events such as demand, fluctuations, price adjustment etc. within legal boundaries (Mikalef & Krogstie, 2018). What remains unclear is how to anonymize and govern captured datasets without bridging legal regulations. Recent studies highlighted the incremental value and importance of diagnosticity, diversity and governance of big data (Ghasemaghaei & Calic, 2019a; Grange et al., 2019; Mikalef & Krogstie, 2018). In sum, data diagnosticity, diversity and governance have relational dependency on one another, but at the same time are beneficial and desirable facets in terms of informational value for organizations and decision-makers.

5 Conclusion

Using simulative, hypothetical, current and historical data to move from not knowing how to solve a problem to knowing how to solve it is data-driven insight through big data analytics. It has been shown to help managers to improve their business strategies and decision-making processes (Ghasemaghaei & Calic, 2019b; Sivarajah et al., 2017). Huge datasets represent collections of multiple observations which comes in various shapes, size, formats, volume and even non-Euclidean objects (i.e., networks- associations between edges and nodes in social media platforms) as illustrated by (Dryden & Hodge, 2018). Unfortunately, most times we often have methods that are not scalable, that is, hardly can manage or filter semi-structured or unstructured data into desirable structure for analytical purpose.

A peculiar area for big data is the messiness, missing and biased nature of the observations. Conventional approaches include the use of statistical inference, that is, sampling strategies to draw conclusions on the general population. Despite enthusiastic claims about the effectiveness of sampling strategies and statistical models, drawing causal inference and relationships requires known datasets, formats and preclusion of needed data. In this case, the effectiveness of advanced statistical approaches deems to be in fact marginal when big data is involved.

Thanks to big data and the internet, firms are now able to obtain vast quantities and types of datasets. Thus, new statistical thinking must arise to help us make sense of non-ideal sampling paradigms and develop mechanisms to enable repeatable, defensible inferential conclusions to be drawn (Olhede & Wolfeb,2018). Dealing with big data equates to limited access to information, disorganize sample, biased and/or missing data. Olhede and Wolfeb (2018) stated that diversity of datasets has no unique space structure which raises questions as to how can firms “generate more heterogeneous yet structured observations that better match network data sets” and “how firms integrate covariates and other non-network information into a consistent framework for inference mechanism”? Analyzing and interpreting data from subjecst with high degree of heterogeneity possess huge challenge for firms (Bühlmann & Meinshausen, 2016). The ability of firms to recognize, analyze, interpret and benefit from heterogeneous datasets is significant for competitive advantage e.g., ability to know what, and when to aggregate, smooth, and average versus when to disaggregate and stratify (Ashley, 2016). Modeling, estimating and correctly interpreting complex and random heterogeneous observations remains an outstanding problem (Buhlmann & van der Geer, 2018). Henceforth, estimation and inference strategies that are designed to manage heterogenous datasets are desperately needed, because absence of such modeling tools means that firms cannot leverage on all the potentials of the obtained datasets.

There are enormous implications for data governance e.g., privacy implications. Traditional principles in dealing with observations from human subjects (e.g., informed consent, confidentiality and anonymity assurance) are under pressure. Moreover, the scale and levels of pervasiveness are less compared to those involve in big data. These has led scientist to develop methods that can calculate meaningful summaries from anonymized data (Olhede & Wolfeb,2018). For instance, rather than informed consent as with classical approach, Aslett et al. (2015) recommended encryption and anonymization schemes to ensure privacy. Apart from technical solutions, there efforts and pressures from international bodies on data governance. For example, European General Data Protection Regulation laid down overarching principles on the usage of personally identifiable information e.g., social media handlers. Individuals can choose to have grant free access to their personal data, individuals right to revoke their information from databases and consequences bridging these principles. Other bodies such as the British Academy & Royal Society (2017) and IEEE (2016) have proposed similar principles. We propose that future research should empirically test the phenomenon of interest to broaden our insights.

References

Appelbaum, D., Kogan, A., Vasarhelyi, M., & Yan, Z. (2017). Impact of business analytics and enterprise systems on managerial accounting. International Journal of Accounting Information Systems, 25, 29-44.

Ashley, E.A. (2016). Towards precision medicine. Nature

Reviews Genetics, 17(9), 507.

Aslett, L.J., Esperança, P. M., & Holmes, C.C. (2015). A review of homomorphic encryption and software tools for encrypted statistical machine learning. arXiv preprint arXiv:1508.06574.

Bakshy, E., Messing, S., & Adamic, L.A. (2015). Exposure to ideologically diverse news and opinion on Facebook. Science, 348(6239), 1130-1132.

Bühlmann, P., & Meinshausen, N. (2015). Magging: maximin aggregation for inhomogeneous large-scale data. Proceedings of the IEEE, 104(1), 126-135.

Bühlmann, P., & van de Geer, S. (2018). Statistics for big data: A perspective. Statistics & Probability Letters, 136, 37-41.

British Academy & Royal Society (2017). Data management and use: Governance in the 21st century. Google Scholar

Cai, L., & Zhu, Y. (2015). The challenges of data quality and data quality assessment in the big data era. Data science journal, 14.

Chen, D.Q., Preston, D.S., & Swink, M. (2015). How the use of big data analytics affects value creation in supply chain management. Journal of Management Information Systems, 32(4), 4-39.

Deka, G.C. (2014). Big data predictive and prescriptive analytics. In Handbook of Research on Cloud Infrastructures for Big Data Analytics (pp. 370-391). IGI Global.

Dryden, I.L., & Hodge, D.J. (2018). Journeys in big data statistics. Statistics & Probability Letters, 136, 121-125.

Ghasemaghaei, M., & Calic, G. (2019a). Can big data improve firm decision quality? The role of data quality and data diagnosticity. Decision Support Systems, 120, 38-49.

Ghasemaghaei, M., & Calic, G. (2019b). Does big data enhance firm innovation competency? The mediating role of data-driven insights. Journal of Business Research, 104, 69-84.

Ghasemaghaei, M., Ebrahimi, S., & Hassanein, K. (2018). Data analytics competency for improving firm decision-making performance. The Journal of Strategic Information Systems, 27(1), 101-113.

Ghasemaghaei, M., Hassanein, K., & Turel, O. (2017). Increasing firm agility through the use of data analytics: The role of fit. Decision Support Systems, 101, 95-105.

Grange, C., Benbasat, I., & Burton-Jones, A. (2019). With a little help from my friends: Cultivating serendipity in online shopping environments. Information & Management, 56(2), 225-235.

Hazen, B.T., Boone, C.A., Ezell, J.D., & Jones-Farmer, L. A. (2014). Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications. International Journal of Production Economics, 154, 72-80.

IEEE (2016). Ethically aligned design, v1

Kwon, O., Lee, N., & Shin, B. (2014). Data quality management, data usage experience and acquisition intention of big data analytics. International journal of information management, 34(3), 387-394.

Marsden, J. R., & Pingry, D. E. (2018). Numerical data quality in IS research and the implications for replication. Decision Support Systems, 115, A1-A7.

Mikalef, P., & Krogstie, J. (2018). Big Data Governance and Dynamic Capabilities: The Moderating effect of Environmental Uncertainty. In Twenty-Second Pacific Asia Conference on Information Systems. Japan

Olhede, S.C., & Wolfeb, P.J. (2018). The future of statistics and data science. Statistics & Probability Letters, 136, 46-50

Sivarajah, U., Kamal, M. M., Irani, Z., & Weerakkody, V. (2017). Critical analysis of Big Data challenges and analytical methods. Journal of Business Research, 70, 263-286.

Sumanth, S. (2019). Discover ideas about Big Data Machine Learning. Retrieved from https://www.pinterest.com/pin/550776229409314191/ (Accessed August 2019)

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. Journal of Management Information Systems, 12(4), 5-33.

Weber, K., Otto, B., & Österle, H. (2009). One Size Does Not Fit All---A Contingency Approach to Data Governance. Journal of Data and Information Quality, 1(1), 1–27. https://doi.org/10.1145/1515693.1515696