Journal of Big Data

RESEARCH Open Access

Exploring big data traits and data quality dimensions for big data analytics application using partial least squares structural equation modelling

Muslihah Wook[*], Nor Asiakin Hasbullah, Norulzahrah Mohd Zainudin, Zam Zarina Abdul Jabar, Suzaimah Ramli, Noor Afiza Mat Razali and Nurhafizah Moziyana Mohd Yusop

*Correspondence: muslihah@upnm.edu.my
Department of Computer Science, Faculty of Defence Science and Technology, National Defence University of Malaysia, Kem Perdana Sungai Besi, 57000 Kuala Lumpur, Malaysia

Abstract

The popularity of big data analytics (BDA) has boosted the interest of organisations into exploiting their large scale data. This technology can become a strategic stimulation for organisations to achieve competitive advantage and sustainable growth. Previous BDA research, however, has focused more on introducing more traits, known as Vs for big data traits, while ignoring the quality of data when examining the application of BDA. Therefore, this study aims to explore the effect of big data traits and data quality dimensions on BDA application. This study has formulated 10 hypotheses that comprised of the relationships of big data traits, accuracy, believability, completeness, timeliness, ease of operation, and BDA application constructs. This study conducted a survey using a questionnaire as a data collection instrument. Then, the partial least squares structural equation modelling technique was used to analyse the hypothesised relationships between the constructs. The findings revealed that big data traits can significantly affect all constructs for data quality dimensions and that the ease of operation construct has a significant effect on BDA application. This study contributes to the literature by bringing new insights to the field of BDA and may serve as a guideline for future researchers and practitioners when studying BDA application.

Keywords: Big data analytics, Big data, Big data traits, Data quality dimensions, Partial least squares structural equation modelling, Survey questionnaire

Introduction

Driven by globalisation and increasing market competitions, various industries have turned to big data analytics (BDA) for its ability to transform enormous raw data into decision-making tools [1]. BDA consists of a set of advanced analytical techniques adapted from related fields, such as artificial intelligence, statistics, and mathematics, which are used to identify trends, detect patterns, and unveil hidden knowledge from a huge amount of data [2]. This technology has been applied in different fields, including finance [3], insurance [4], and cyber security [5], to name a few. The emergence of BDA can be linked to the inability of traditional database management tools to handle structured and unstructured data simultaneously [6]. Structured data refers to data that have a scheme, metadata, rules, and constraints to follows, whilst unstructured data have no structure at all or unknown structure to follow [7]. These types of data are collected or received from diverse platforms, such as network sensors, social media, and the Internet of Things.

Although it is vital to exploit structured and unstructured data for BDA, they are usually incomplete, inaccurate, inconsistent, and vague or ambiguous, which could lead to false decisions [8–11]. Salih et al. [12] and Wamba et al. [13] have highlighted the lack of data quality mechanisms being applied in BDA prior to data usage. Several studies have considered the potential of data quality for BDA application [14–18], yet, specific questions about what drives the dimensions of data quality remain unanswered. Nevertheless, studies on data quality and BDA are still underway and have not reach a good level of maturity [7]. Thus, there is an urgent need to conduct in-depth study on data quality to determine the most important dimensions for BDA application.

Several theories or models for understanding data quality problems have been suggested, such as resource-based theory (RBT), organisational learning theory (OLT), firm performance (FPER), and data quality framework (DQF). However, these theories or models do not fit into BDA application since they concentrate primarily on service quality as opposed to data quality [19]. Moreover, most studies related to BDA are focused on the perspective held at the organisational or firm level [8, 10, 20, 21] and studies focusing on the individual perspective are lacking. Since academics are encouraged to participate in research on pedagogical support for teaching about BDA [22], this study has determined that university students can represent the perspectives at the individual level. Students were chosen because it is crucial to prepare and expose them to BDA, especially in the mandatory setting [23].

Meanwhile, numerous traits have been studied to explain the characteristics of big data, such as 3Vs [24], 4Vs [25], 5Vs [26, 27], 7Vs [28], 9Vs [29], 10Vs [30], 10Bigs [31], and 17Vs [32]. These attempts to assign the maximum number of characteristics to big data show the lack of uniform consensus regarding the core of big data characteristics [33]. Although big data characteristics and data quality are viewed as distinct domains, several studies have found that these two domains are interconnected and closely related [9, 14, 17]. A better understanding of the core characteristics of big data and the dimensions of data quality is needed. Hence, this study seeks to expand the knowledge on big data characteristics, hereafter known as big data traits (BDT) and data quality dimensions (DQD), as well as to explore how they could affect the application of BDA.

Literature review

Big data and analytics are two different fields that are widely used to exploit the exponential growth of data in recent years. The term 'big data' represents a large volume of data, while the term 'analytics' indicates the application of mathematical and statistical tools on a collection of data [34]. These two terms have been merged into 'big data analytics' to represent various advanced digital techniques that are formulated to identify hidden patterns of information within gigantic data sets [35, 36]. Scholars have suggested varying definitions for BDA. For instance, Verma et al. [23] defined BDA as a suite of data management and analytical techniques for handling complex data sets, which in turn lead to a better understanding of the underlying process. Faroukhi et al. [37] defined BDA as a process of analysing raw data in order to obtain information that is understandable to humans, which are hard to observe using direct analysis. Davenport [38] simply defined BDA as a "focus on very large, unstructured and fast moving data".

Nowadays, BDA application has helped numerous organisations improve their performance because it can handle problems instantly and assist organisations in making better and smarter decisions [35, 39]. The advantages of BDA application for organisational performance have been proven by numerous studies. For instance, Mikalef et al. [20] found four alternative solutions surrounding BDA that can lead to higher performance, whereby different combinations of BDA resources either play a greater or lesser importance to organisational performance. Similarly, Wamba et al. [40] applied the RBT and sociomaterialism theory to examine organisational performance. Their empirical work showed that the hierarchical BDA has both direct and indirect impacts on organisational performance. Based on this same set of views, Wamba et al. [13] highlighted the importance of capturing the quality dimensions of BDA. Their findings proved the existence of a significant relationship between the quality of data in BDA and organisational performance.

Some scholars perceive data quality as equivalent to information quality [41–44]. Data quality generally refers to the degree to which the data are fit for use [45]. Meanwhile, the concept of information quality is defined as how well the information supports the task [46]. Haryadi et al. [14] asserted that data quality is focused on data that have not been analysed, while information quality is focused on the analysis that has been done on the data. This study, however, opines that data quality should focus on the wellness and appropriateness of data, which encompasses either before or after it has been analysed, in which it should meet the requirements of organisations [12].

The notion of quality represents a multidimensional construct, whereby it is essential to combine its dimensions and express them in a solid structure [46]. Initially, Wang and Strong [45] used factor analysis to identify DQD and found 179 dimensions that were eventually reduced to 20. Then, they organised these dimensions into four primary categories, namely intrinsic, contextual, representational, and accessibility. The intrinsic category denotes datasets that have quality in their own right, while the contextual category highlights the requirement of the task that data quality must be considered within the context. The representational category describes data quality in relation to the presentation of the data, and the accessibility category emphasises on the importance of computer systems that provide access to data [18]. Each category has several dimensions that are used as specific data quality measurements. For instance, accuracy and objectivity are the dimensions in the intrinsic category, while relevance and timeliness are the dimensions in the contextual category. Interpretability and understandability are the dimensions in the representational category, and access security and ease of operations are the dimensions in the accessibility category. Table 1 presents all DQD according to their categories.

Various studies have been conducted to analyse the relationships between DQD and BDA application. For instance, Côrte-Real et al. [8] analysed the direct and indirect effects of DQD on BDA capabilities in a multi-regional survey (European and

Table 1 DQD and their categories [17]

DQD	Data quality categories
Accuracy	Intrinsic
Objectivity	
Believability	
Reputation	
Value-added	Contextual
Relevancy	
Timeliness	
Completeness	
Appropriate amount of data	
Interpretability	Representational
Understandability	
Concise representation	
Consistent representation	
Accessibility	Accessibility
Access security	
Ease of operations	

American firms). Their findings showed that the DQD, primarily completeness, accuracy, and currency, have significant effects on BDA capabilities when process complexity was low. Thus, these authors have demonstrated the emergent need for firms to have effective data quality mechanisms to be able to derive sufficient value from BDA application. Ghasemaghaei and Calic [47] used OLT and the DQD compiled by Wang and Strong [45] to explain the effect of BDA on data quality categories. They found that while many organisations have invested in BDA application, they need to pay more attention to the quality of their data in order to enhance the quality of the solutions. Meanwhile, Ji-fan Ren et al. [48] examined the quality dynamics (system quality and information quality) in BDA using business value and FPER theories. Their study revealed that system quality can enhance information quality, which in turn, would affect organisational values and performance in the BDA environment. While these studies offer insights into the relationship between DQD and BDA, they have not highlighted the critical DQD that could impact BDA application.

DQD are also associated with the characteristics of big data, which are commonly known as big data traits (BDT). The BDT were originally defined by 3Vs (volume, velocity, and variety) [24]. These traits have been extended over the years, which include 4Vs (volume, velocity, variety, and value) [25], 5Vs (volume, velocity, variety, value, and veracity) [26, 27], 7Vs (volume, velocity, variety, veracity, validity, volatility, and value) [28], 9Vs (veracity, variety, velocity, volume, validity, variability, volatility, visualisation, and value) [29], 10Vs (volume, value, velocity, veracity, viscosity, variability, volatility, viability, validity, and variety) [30], 10Bigs (big volume, big velocity, big variety, big veracity, big intelligence, big infrastructure, big service, big value, and big market) [31], and 17Vs (volume, velocity, value, variety, veracity, validity, volatility, visualisation, virality, viscosity, variability, venue, vocabulary, vagueness, verbosity, voluntariness, and versatility) [32].

Several studies have investigated the influence of BDT on DQD. Noorwali et al. [49] argued that there is a lack of scientific understanding of the general and specific requirements of BDT and DQD. They suggested that a more systematic analysis for both BDT and DQD is essential for reducing the number of missing quality requirements while accounting for BDT. Likewise, Lakshen et al. [50] argued on the various technical challenges that must be addressed before the potential of BDT and DQD can be fully realised. Haryadi et al. [14] confirmed that BDT and DQD are the central issues for implementing BDA.

Conceptual model and hypotheses

Based on the literature review in the previous section, this study proposes a new model for BDA application based on the integration of BDT and DQD, as depicted in Fig. 1. It should be noted that various applications can have different requirements, as not all dimensions and constructs are always applicable. Nevertheless, most studies on BDA application are focused on the organisation or firm levels and not on the individual level. Hence, this study is based on an individual's perception.

Big data traits

According to Sun [31], various Vs are used to define BDT, while conventional data quality is defined by a number of DQD [17]. Hence, this study considered BDT as a single construct because different Vs are overlapping with the DQD. DQD categories that are generally accepted and frequently used in the application of BDA were also included in this study, which

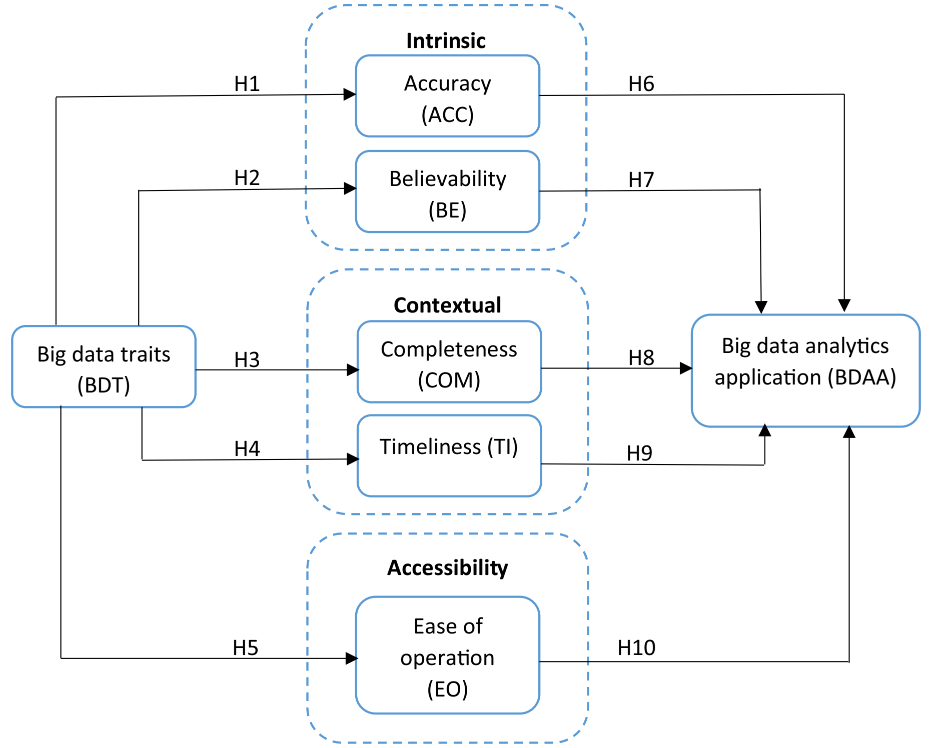

Fig. 1 The proposed conceptual model for BDA application

were the intrinsic, contextual, and accessibility categories. The intrinsic category was chosen because of the importance of data correctness in BDA application, which is composed of two constructs, namely, accuracy and believability. Meanwhile, the contextual category was chosen because the application of BDA commonly depends on the context in which the data are used. This study considered two constructs in the contextual category, namely, completeness and timeliness. Finally, the accessibility category was chosen because the computer system needs to facilitate the accessing and storing of data in BDA application. Thus, the ease of operation is considered as a construct in the accessibility category for this study. The significant influence of BDT on the constructs of DQD, namely, accuracy, believability, completeness, timeliness, and ease of operation was explored through the following hypotheses:

H1: Big data traits have a significant influence on accuracy.

H2: Big data traits have a significant influence on believability.

H3: Big data traits have a significant influence on completeness.

H4: Big data traits have a significant influence on timeliness.

H5: Big data traits have a significant influence on ease of operation.

Accuracy

Accuracy means that the data must depict facts accurately and the data must come from a valid source [45, 51]. The effective use of BDA relies on the accuracy of data, which is necessary to produce reliable information [8]. As higher data accuracy may facilitate the routines and activities of BDA, this study proposes that accuracy is included as an enabler in BDA application. Hence, this study proposes the following hypothesis:

H6: Accuracy has a significant influence on big data analytics application.

Believability

Believability represents the degree of which the data is considered valid and reliable [44]. There are concerns regarding the credibility of BDA findings due to insufficient insight into the trustworthiness of the data source [52]. Believability of data sources might be difficult to notice, as people may alter facts or even publish false information. Therefore, data sources need to be treated as believable in BDA application. Hence, this study proposes the following hypothesis:

H7: Believability has a significant influence on big data analytics application.

Completeness

Completeness refers to the degree of which there is no lack of data and that the data are largely appropriate for the task at hand [45]. It also refers to the validity of the values of all components in the data [53]. As big data sources are rather large and the architectures are complicated, the completeness of data is crucial to avoid errors and inconsistencies in the outcome of BDA application. Hence, this study proposes the following hypothesis:

H8: Completeness has a significant influence on big data analytics application.

Timeliness

Timeliness refers to the degree of which data from the appropriate point in time reflects truth [50]. Timeliness is identified as one of the most significant dimensions of data quality, since making decisions based on outdated data will ultimately lead to incorrect insights [54]. Additionally, the more rapidly data are being generated and processed, the better time the data will be used in BDA application [17]. Hence, this study proposes the following hypothesis:

H9: Timeliness has a significant influence on big data analytics application.

Ease of operation

Ease of operation refers to the degree of which data can be easily merged, changed, updated, downloaded or uploaded, aggregated, reproduced, integrated, customised, and manipulated, as well as can be used for multiple purposes [45]. Users will undeniably face challenges and complexity to utilise BDA based on the technical approaches used for handling this technology [35]. If the BDA application is relatively easily to operate, the user would be willing to use it in a long-term. Hence, this study proposes the following hypothesis:

H10: Ease of operation has a significant influence on big data analytics application.

Research methodology

The methodological procedures in this study were conducted in two phases, namely, research instrument, and data collection and analysis. Figure 2 shows the methodology in sequence.

Research instrument

This study used a survey questionnaire with two sections to explore the hypothesised relationships in the proposed conceptual model. The first section included questions related to the respondents' profiles, such as gender, year of study, and area of study, while the second section contained measurement of constructs with 28 indicators. These constructs were BDT, accuracy, believability, completeness, timeliness, ease of operation, and BDA application. The indicators to measure BDT (velocity, veracity, value, and variability) were self-developed based on the definitions proposed by Arockia et al. [32]. The accuracy, believability, completeness, timeliness, and ease of operation constructs, each with four indicators, were adapted from [8, 47], and [48]. BDA application, with four indicators, was adapted from [23] and [55]. All indicators have been measured using

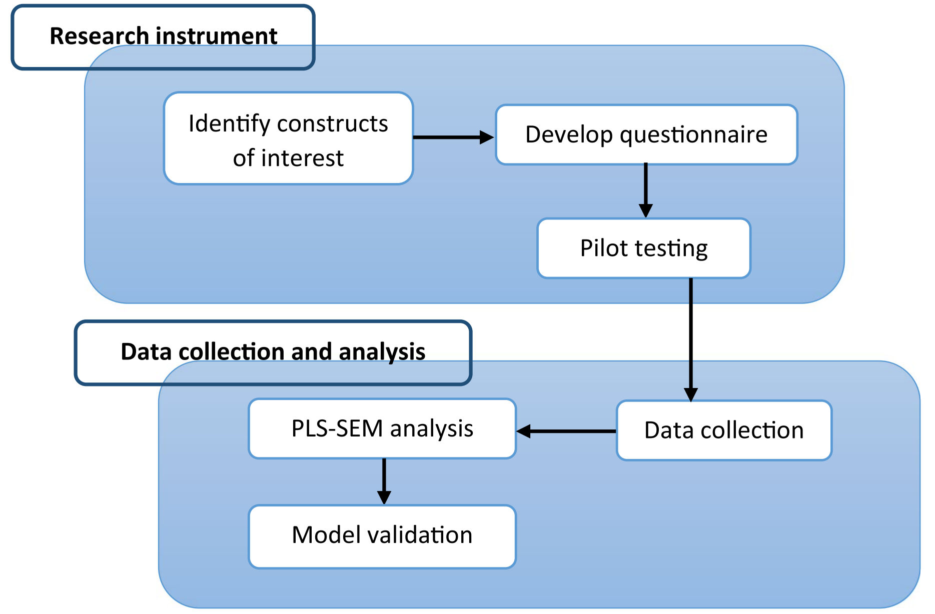

Fig. 2 Methodological procedures

the 7-point Likert scale, ranging from 1 (strongly disagree) to 7 (strongly agree). This questionnaire was pretested among academics and several items have been reworded to improve the clarity of the questions. The questionnaire was then used in a pilot test to confirm the reliability of all shortlisted constructs. This test involved 30 respondents, and the Cronbach's alpha values for all seven constructs on the reliability scale were found to be appropriate and acceptable.

Data collection and analysis

This study used random sampling to select respondents who have knowledge on BDA. As a preliminary study, 200 survey invitations were sent to Computer Science students at the National Defence University of Malaysia. These students were chosen because of the knowledge they had gained during the Big Data Analytics or Data Mining course that they attended previously. Data were collected through a web survey, which was conducted from July till August 2020. A total of 108 complete responses were received, resulting in 54% response rate. There were 84 male (77.78%) and 24 female (22.22%) respondents involved in this study. Most of the respondents were in their second year of study (52.78%) and in the area of artificial intelligence (50.93%). The key profiles of these respondents are shown in Table 2.

Subsequently, the partial least squares structural equation modelling (PLS-SEM) was applied to analyse the survey-based crossectional data, since this technique is able to explain the variance in key target constructs [56]. This technique amalgamates the concepts of factor analysis and multiple regression in order to validate the measurement instruments and test the research hypotheses. Since PLS-SEM is a modest and practical technique to create rigour in a complex modeling [57], this study had also utilised this technique for analysing and validating the complex hypothesised relationships of the proposed model.

Table 2 Respondents' profiles

Description	Frequency	Percentage
Gender		
Male	84	77.78
Female	24	22.22
Total	108	100
Year of study		
Year 1	26	24.07
Year 2	57	52.78
Year 3	15	13.89
Graduated	10	9.26
Total	108	100
Area of study		
Computer security	10	9.26
Computer science	43	39.81
Artificial intelligence	55	50.93
Total	108	100

Analysis and results

The SmartPLS 3.2 package was used to perform the PLS-SEM analysis. Evaluation of the analysis began with the measurement model, followed by the structural model.

Measurement model

A measurement model was used to assess the reliability and validity of the constructs. The standard steps for assessing a measurement model are convergent validity and discriminant validity. Convergent validity was analysed by calculating the factor loading of the indicators, composite reliability (CR), and average variance extracted (AVE) [58]. The convergent validity results in Table 3 show that the factor loadings for all indicators are higher than 0.708, as suggested by Hair et al. [59], with the elimination of three indicators (AC3, BE2, and EO4) from the original 28 indicators. Meanwhile, the CR values ranged from 0.822 to 0.917, which exceeded the suggested value of greater than 0.7 [59]. An adequate AVE is 0.50 or greater, meaning that at least 50% of the variance of the constructs can be explained by its indicators [56]. As shown in Table 3, all AVE values range from 0.536 to 0.786, indicating that convergent validity of the measurement model is achieved.

Once the convergent validity has been successfully established, the discriminant validity was examined using the Fornell-Larcker criterion. The square root of AVE should be greater than the correlations among each construct [60]. Table 4 demonstrates that the square root of AVEs are greater in all cases than the off-diagonal elements in their corresponding row and column. Therefore, discriminant validity has been achieved.

Structural model

The structural model was used to examine the magnitude of the relationships among the constructs. The goodness of fit of the structural model can be assessed by examining the R^2 measure (the coefficient of determination) and the significance level of the path

Table 3 Convergent validity results

No	Constructs	Indicators	Factor Loadings	CR	AVE
1	Big data traits (BDT)	V1	0.725	0.822	0.536
		V2	0.765		
		V3	0.726		
		V4	0.712		
2	Accuracy (AC)	AC1	0.763	0.861	0.675
		AC2	0.907		
		AC4	0.786		
3	Believability (BE)	BE1	0.823	0.869	0.688
		BE3	0.866		
		BE4	0.798		
4	Completeness (COM)	COM1	0.786	0.881	0.649
		COM2	0.831		
		COM3	0.771		
		COM4	0.832		
5	Timeliness (TI)	TI1	0.805	0.893	0.677
		TI2	0.758		
		TI3	0.854		
		TI4	0.869		
6	Ease of operation (EO)	EO1	0.844	0.917	0.786
		EO2	0.905		
		EO3	0.910		
7	Big data analytics application (BDAA)	BDA1	0.867	0.901	0.695
		BDA2	0.830		
		BDA3	0.869		
		BDA4	0.764		

Table 4 Discriminant validity results

	AC	BE	BDAA	BDT	COM	EO	TI
AC	0.821						
BE	0.748	0.829					
BDAA	0.558	0.595	0.834				
BDT	0.653	0.674	0.530	0.732			
COM	0.627	0.682	0.520	0.576	0.806		
EO	0.564	0.593	0.565	0.561	0.547	0.887	
TI	0.587	0.666	0.542	0.534	0.757	0.656	0.823

coefficients (β values) [56]. The results of the research model were satisfactory, demonstrating the R^2 value for BDA application at 0.444, which suggested that 44.4% of the variance in BDA application can be explained by DQD. Furthermore, the R^2 values for constructs of accuracy (42.6%), believability (45.5%), and completeness (33.2%) were also satisfactory, except for timeliness (26.8%) and ease of operation (31.5%) that were moderately explained by BDT. Figure 3 illustrates the results of the R^2 values from the Smart-PLS 3.2 software.

The path coefficents of the structural model were calculated using bootstrap analysis (resampling = 5000) to assess their statistical significance. Table 5 shows the results of

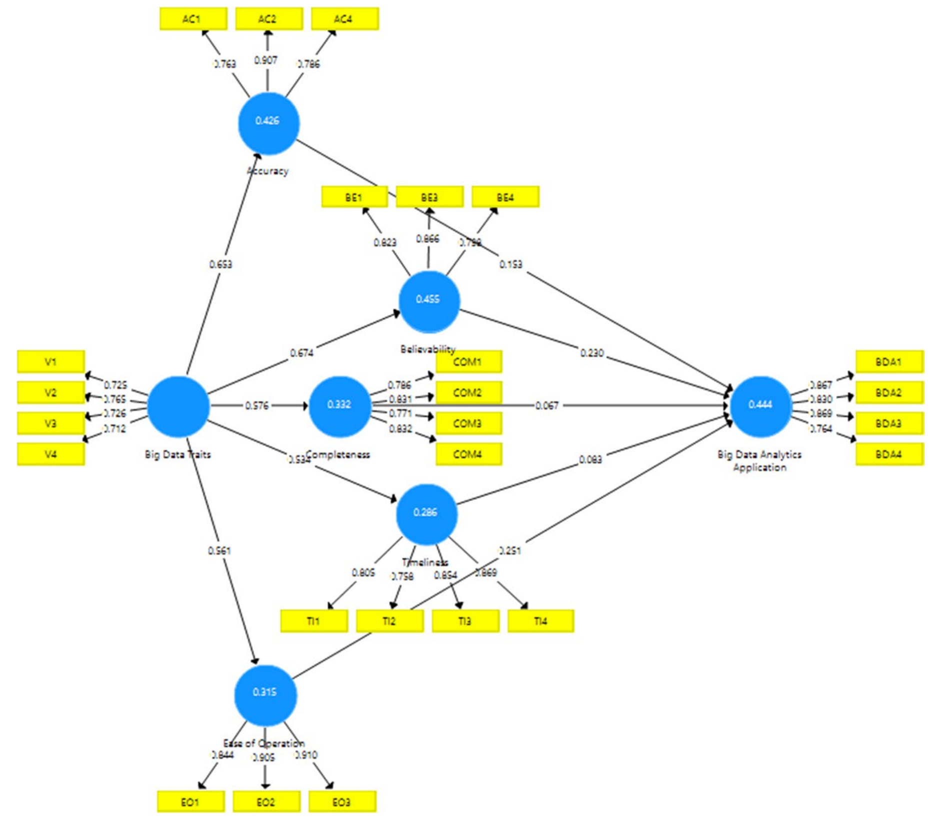

Fig. 3 The results of R^2 values

Table 5 The analysis results of the structural model

Hypotheses	Constructs	Path coefficient (β)	*t*-value	*p*-value	Results
H1	BDT → AC	0.653	15.507	0.000	Significant
H2	BDT → BE	0.674	17.751	0.000	Significant
H3	BDT → COM	0.576	9.796	0.000	Significant
H4	BDT → TI	0.534	8.514	0.000	Significant
H5	BDT → EO	0.561	7.142	0.000	Significant
H6	AC → BDAA	0.153	1.298	0.097	Non-significant
H7	BE → BDAA	0.230	1.485	0.069	Non-significant
H8	COM → BDAA	0.067	0.476	0.317	Non-significant
H9	TI → BDAA	0.083	0.581	0.281	Non-significant
H10	EO → BDAA	0.251	1.922	0.028	Significant

Critical values: $t > 1.645$; $p < 0.05$

the path coefficients and their level of significance. An analysis of *t*-value and *p*-value has shown that six hypotheses were significant, namely, H1, H2, H3, H4, H5, and H10. Overall, H1 to H5 were the influence of BDT on DQD, whereas, only one hypothesis of DQD, H10, was identified as significant for evaluating the influence of ease of operation towards BDA application. The results have also shown that accuracy, believability, completeness, and timeliness had no significant effect on BDA application. Thus, H6, H7, H8, and H9 were rejected.

Discussion

The present study has explored the BDT and DQD constructs for BDA application. The findings showed that the accessibility of DQD (ease of operation) can significantly influence BDA application. This result showed that the ease of obtaining data plays an important role in providing users with an effective access to reduce the digital divide in BDA application endeavour. This result is corroborated by the findings by Zhang et al. [61], who considered the ease of functional properties would ensure the quality of BDA application. Janssen et al. [9] similarly proposed that the easier it is to operate BDA, the more application systems would be integrated and are sufficient for handling this technology.

Akter et al. [57] found significant influence of DQD (completeness, accuracy, format, and currency) on BDA application. On the other hand, the results of this study showed that accuracy, believability, completeness, and timeliness have no significant influence on the decision to apply BDA. These results were unexpected. These outcomes could be because the respondents were novice users, whom assumed the availability of technical teams to solve any accuracy, believability, completeness, and timeliness problems in BDA application.

Meanwhile, the four indicators of BDT (velocity, veracity, value, and variability) have shown significantly high impact on all constructs of DQD (accuracy, believability, completeness, timeliness, and ease of operation). These findings are in agreement with the results obtained by Wahyudi et al. [17], whereby high correlation was found between BDT, and timeliness and ease of operation. The significant influence of BDT on DQD showed interesting results, which demonstrated how users recognise the importance of BDT for assessing quality assessment results. This observation is in agreement with Taleb et al. [62], who claimed that BDT could enforce quality evaluation management to achieve quality improvements. The findings also showed that while many researchers have proposed numerous BDT, in this context, velocity, veracity, value, and variability are more critical for assessing data quality in BDA application.

Conclusion

This study has proposed the practical implications based on perspectives at the individual level. Individual perspectives are imperative since the resistance to use technology commonly originates from this level of users. Hence, the results of this study may be beneficial for organisations that have not yet agreed to implement BDA. They could use the results to have a sense of the possibilities from embracing this technology. This study has also shown the theoretical implications based on the incorporation of BDT as a single construct and DQD as an underpinning theory for the development of a new BDA application model. This study is the first to investigate the influence of BDT and DQD towards BDA application by individual level users.

Several limitations apply to the interpretation of the results in this study. First, the intrinsic and contextual data quality categories are inadequate to specify the DQD included in the proposed model. Future studies may include other DQD, such as objectivity and reputation to represent the intrinsic category. Meanwhile, value-added, relevancy, and appropriate amount of data can be used for measuring the contextual category. Second, the chosen undergraduate students who have knowledge on BDA were insufficient to generalise the individual level perceptions towards BDA application. Hence, future studies could include more experienced respondents, such as lecturers or practitioners. Third, although the sample size was statistically sufficient, a larger sample may be useful to reinforce the results of this study. Finally, although this study has attempted to bridge the gaps between BDT and DQD, future studies are encouraged to explore other constructs for better understanding of BDA application. For instance, future studies could explore the role of security and privacy concerns in BDA application since data protection is becoming more crucial due to recent big open data initiatives. Therefore, a novel BDA application model that can address security and privacy concerns may be worth exploring. Overall, the findings of this study have contributed to the body of knowledge in the BDA area and offered greater insights for BDA application initiators.

Abbreviations

BDA: Big data analytics; DQD: Data quality dimensions; BDT: Big data traits; PLS-SEM: Partial least squares structural equation modelling.

Acknowledgements

The authors would like to thank the Centre for Research and Innovation Management, National Defence University of Malaysia for supporting this research (UPNM/2018/GPJP/2/TK/5).

Authors' contributions

MW conceptualised the idea and led the research. NAH and NMZ wrote the literature review and designed the proposed model. MW and ZZAJ contributed to data collection and analysis processes. SR and NAMR drafted and proofread the manuscript. MW and NMMY reviewed the manuscript. All authors read and approved the final manuscript.

Funding

The research is funded under the Short Term Grant program provided by the Centre for Research and Innovation Management, National Defence University of Malaysia, 2018—2020 (Grant code: UPNM/2018/GPJP/2/TK/5).

Availability of data and materials

The datasets are available from the corresponding author upon reasonable request.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 19 December 2020 Accepted: 13 March 2021
Published online: 23 March 2021

References

1. Seyedan M, Mafakheri F. Predictive big data analytics for supply chain demand forecasting: methods, applications, and research opportunities. J Big Data. 2020. https://doi.org/10.1186/s40537-020-00329-2.
2. Saggi MK, Jain S. A survey towards an integration of big data analytics to big insights for value-creation. Inf Process Manag. 2018;54(5):758–90. https://doi.org/10.1016/j.ipm.2018.01.010.
3. Hasan MM, Popp J, Oláh J. Current landscape and influence of big data on finance. J Big Data. 2020. https://doi.org/10.1186/s40537-020-00291-z.
4. Arumugam S, Bhargavi R. A survey on driving behavior analysis in usage based insurance using big data. J Big Data. 2019. https://doi.org/10.1186/s40537-019-0249-5.

5. Obitade PO. Big data analytics: a link between knowledge management capabilities and superior cyber protection. J Big Data. 2019. https://doi.org/10.1186/s40537-019-0229-9.
6. Boukhalfa A, Hmina N, Chaoui H. Survey on security monitoring and intrusion detection in the big data environment. Int J Adv Trends ComputSciEng. 2020;9(4):6175–9.
7. Taleb I, Serhani MA, Dssouli R. Big data quality assessment model for unstructured data. Proc IntConf 2018 13th InnovInfTechnol IIT. 2018;2019:69–74.
8. Côrte-Real N, Ruivo P, Oliveira T. Leveraging internet of things and big data analytics initiatives in European and American firms: Is data quality a way to extract business value? InfManag. 2020;57(1):103141. https://doi.org/10.1016/j.im.2019.01.003.
9. Janssen M, van der Voort H, Wahyudi A. Factors influencing big data decision-making quality. J Bus Res. 2017;70:338–45. https://doi.org/10.1016/j.jbusres.2016.08.007.
10. Ghasemaghaei M. Are firms ready to use big data analytics to create value? The role of structural and psychological readiness. EnterpInfSyst. 2019;13(5):650–74. https://doi.org/10.1080/17517575.2019.1576228.
11 Hariri RH, Fredericks EM, Bowers KM. Uncertainty in big data analytics: survey, opportunities, and challenges. J Big Data. 2019. https://doi.org/10.1186/s40537-019-0206-3.
12. Salih FI, Ismail SA, Hamed MM, MohdYusop O, Azmi A, MohdAzmi NF. Data quality issues in big data: a review. AdvIntellSystComput. 2019;843:105–16.
13. FossoWamba S, Akter S, de Bourmont M. Quality dominant logic in big data analytics and firm performance. Bus Process Manag J. 2019;25(3):512–32.
14. Haryadi AF, Hulstijn J, Wahyudi A, Van Der Voort H, Janssen M. Antecedents of big data quality: an empirical examination in financial service organizations. Proc 2016 IEEE IntConf Big Data. 2016;2016:116–21.
15. Janssen M, Konopnicki D, Snowdon JL, Ojo A. Driving public sector innovation using big and open linked data (BOLD). InfSyst Front. 2017;19(2):189–95.
16. Merino J, Caballero I, Rivas B, Serrano M, Piattini M. A data quality in use model for big data. FuturGenerComputSyst. 2016;63:123–30.
17 Wahyudi A, Farhani A, Janssen M. Relating big data and data quality in financial service organizations. Lect Notes Comput Sci. 2018. https://doi.org/10.1007/978-3-030-02131-3_45.
18. Wahyudi A, Kuk G, Janssen M. A process pattern model for tackling and improving big data quality. InfSyst Front. 2018;20(3):457–69.
19. Panahy PHS, Sidi F, Affendey LS, Jabar MA, Ibrahim H, Mustapha A. A framework to construct data quality dimensions relationships. Indian J SciTechnol. 2013;6(5):4422–31.
20. Mikalef P, Boura M, Lekakos G, Krogstie J. Big data analytics and firm performance: findings from a mixed-method approach. J Bus Res. 2019;98(February):261–76.
21 Wahdain EA, Baharudin AS, Ahmad MN. Big data analytics in the Malaysian public sector: the determinants of value creation. In: Saeed F, Gazem N, Mohammed F, Busalim A, editors. Advances in intelligent systems and computing, vol. 843. Cham: Springer International Publishing; 2019. p. 139–50.
22. Côrte-Real N, Oliveira T, Ruivo P. Assessing business value of big data analytics in European firms. J Bus Res. 2017;70:379–90. https://doi.org/10.1016/j.jbusres.2016.08.011.
23. Verma S, Bhattacharyya SS, Kumar S. An extension of the technology acceptance model in the big data analytics system implementation environment. Inf Process Manag. 2018;54(5):791–806. https://doi.org/10.1016/j.ipm.2018.01.004.
24. Laney D. 3D data management: controlling data volume, velocity, and variety. Application Delivery Strategies. 2001. https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf. Accessed 25 Sept 2020.
25. Thiyagarajan VS, Venkatachalapathy K. Isolating values from big data with the help of four V'S. Int J Res EngTechnol. 2015;04(01):132–5.
26. JasimHadi H, Hameed Shnain A, Hadishaheed S, Haji AA. Big data and five V'S characteristics. Int J Adv Electron ComputSci. 2015;2:2393–835.
27 Ishwarappa AJ. A brief introduction on big data 5Vs characteristics and hadoop technology. Procedia Comput Sci. 2015;48:319–24.
28. Khan MA, Uddin MF, Gupta N. Seven V's of big data: Understanding big data to extract value. In: Proceedings of the 2014 Zone 1 Conference of the American Society for Engineering Education. IEEE; 2014. p. 1–5.
29. Owais SS, Hussein NS. Extract five categories CPIVW from the 9V's characteristics of the big data. Int J AdvComputSciAppl. 2016;7(3):254–8.
30. Khan N, Alsaqer M, Shah H, Badsha G, Abbasi AA, Salehian S. The 10 Vs, issues and challenges of big data. In: Proceedings of the 2018 International Conference on Big Data and Education. ACM; 2018. p. 52–56.
31. Sun Z. 10 Bigs : Big data and its ten big characteristics. In: BAIS No. 17010, PNG University of Technology. 2018; p. 1–10.
32. Arockia PS, Varnekha SS, Veneshia KA. The 17 V's of big data. Int Res J Eng Technol. 2017;4(9):3–6.
33. Wook M, Jabar ZZA, Halim MH, Razali NAM, Ramli S, Hasbullah NA, et al. Big data analytics application model based on data quality dimensions and big data traits in public sector. Int J Adv Trends ComputSciEng. 2020;9(2):1247–56.
34. Sanders NR. How to use big data to drive your supply chain. Calif Manage Rev. 2016;58(3):26–48.
35. Alswedani S, Saleh M. Big data analytics: importance, challenges, categories, techniques, and tools. Int J Adv Trends ComputSciEng. 2020;9(4):5384–92.
36 Favaretto M, De Clercq E, Elger BS. Big data and discrimination: perils, promises and solutions A systematic review. J Big Data. 2019. https://doi.org/10.1186/s40537-019-0177-4.
37 Faroukhi AZ, El Alaoui I, Gahi Y, Amine A. Big data monetization throughout big data value chain: a comprehensive review. J Big Data. 2020. https://doi.org/10.1186/s40537-019-0281-5.
38. Davenport T. Big data at work: dispelling the myths, uncovering the opportunities. Harvard: Harvard Business Review Press; 2014.

39 Shabbir MQ, Gardezi SBW. Application of big data analytics and organizational performance: the mediating role of knowledge management practices. J Big Data. 2020. https://doi.org/10.1186/s40537-020-00317-6.

40 Wamba SF, Gunasekaran A, Akter S, Ren SJF, Dubey R, Childe SJ. Big data analytics and firm performance: effects of dynamic capabilities. J Bus Res. 2017;70:356–65. https://doi.org/10.1016/j.jbusres.2016.08.009.

41. Boritz JE. IS practitioners' views on core concepts of information integrity. Int J Account InfSyst. 2005;6(4):260–79.

42. Knight SA, Burn J. Developing a framework for assessing information quality on the World Wide Web. Informing Sci. 2005;8:159–72.

43. Madnick SE, Wang RY, Lee YW, Zhu H. Overview and framework for data and information quality research. J Data InfQual. 2009;1(1):1–22.

44. Pipino LL, Lee YW, Wang RY. Data quality assessment. Commun ACM. 2002;45(4):211–8.

45. Wang RY, Strong DM. Beyond accuracy: what data quality means to data consumers. J ManagInfSyst. 1996;12(4):5–33.

46 Todoran IG, Lecornu L, Khenchaf A, Le Caillec JM. A methodology to evaluate important dimensions of information quality in systems. J Data Inf Qual. 2015. https://doi.org/10.1145/2744205.

47. Ghasemaghaei M, Calic G. Can big data improve firm decision quality? The role of data quality and data diagnosticity. Decis Support Syst. 2018;2019(120):38–49. https://doi.org/10.1016/j.dss.2019.03.008.

48. Ji-fan Ren S, FossoWamba S, Akter S, Dubey R, Childe SJ. Modelling quality dynamics, business value and firm performance in a big data analytics environment. Int J Prod Res. 2017;55(17):5011–26.

49. Noorwali I, Arruda D, Madhavji NH. Understanding quality requirements in the context of big data systems. In: Proceedings of the 2nd International Workshop on Big Data Software Engineering. ACM; 2016. p. 76–79.

50. Lakshen GA, Vraneš S, Janev V. Big data and quality: A literature review. In: 2016 24th Telecommunications Forum. IEEE; 2016. p. 1–4.

51. Taleb I, Dssouli R, Serhani MA. Big data pre-processing: A quality framework. In: 2015 IEEE International Congress on Big Data. IEEE; 2015. p. 191–198.

52. Toivonen M. Big data quality challenges in the context of business analytics. https://helda.helsinki.fi/handle/10138/156666. Accessed 12 Aug 2020.

53. Cai L, Zhu Y. The challenges of data quality and data quality assessment in the big data era. Data Sci J. 2015;14:1–10.

54. Merino J, Xie X, Parlikad AK, Lewis I, McFarlane D. Impact of data quality in real-time big data systems. CEUR Workshop Proc. 2020;2716:73–86.

55. Yadegaridehkordi E, Nilashi M, Nasir MHNBM, Ibrahim O. Predicting determinants of hotel success and development using Structural Equation Modelling (SEM)-ANFIS method. Tour Manag. 2018;2018(66):364–86. https://doi.org/10.1016/j.tourman.2017.11.012.

56. Hair JF, Risher JJ, Sarstedt M, Ringle CM. When to use and how to report the results of PLS-SEM. Eur Bus Rev. 2019;31(1):2–24.

57. Akter S, FossoWamba S, Dewan S. Why PLS-SEM is suitable for complex modelling? An empirical illustration in big data analytics quality. Prod Plan Control. 2017;28(11–12):1011–21.

58. Haneem F, Kama N, Taskin N, Pauleen D, Abu Bakar NA. Determinants of master data management adoption by local government organizations: an empirical study. Int J Inf Manage. 2018;2019(45):25–43. https://doi.org/10.1016/j.ijinfomgt.2018.10.007.

59. Hair JF, Howard MC, Nitzl C. Assessing measurement model quality in PLS-SEM using confirmatory composite analysis. J Bus Res. 2019;2020(109):101–10. https://doi.org/10.1016/j.jbusres.2019.11.069.

60 Fornell C, Larcker DF. Evaluating structural equation models with unobservable variables and measurement error. J Market Res. 1981. https://doi.org/10.1177/002224378101800104.

61. Zhang P, Zhou X, Li W, Gao J. A survey on quality assurance techniques for big data applications. In: 2017 IEEE Third International Conference on Big Data Computing Service and Applications. IEEE; 2017. p. 313–319.

62. Taleb I, Serhani MA, Dssouli R. Big data quality: A survey. In: 2018 IEEE International Congress on Big Data. IEEE; 2018. p. 166–173.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.