Resolving the Misconceptions on Big Data Analytics Implementation through Government Research Institute in Malaysia

Mohammad Fikry Abdullah[1], Mardhiah Ibrahim[1] and Harlisa Zulkifli[2]

[1]*Water Resources and Climate Change Research Centre, National Hydraulic Research Institute of Malaysia (NAHRIM), Seri Kembangan, Malaysia*

[2]*Information Management Division, National Hydraulic Research Institute of Malaysia (NAHRIM), Seri Kembangan, Malaysia*

Keywords: Big Data Analytics, Data Management, Misconception, Advanced Analytics.

Abstract: Evolution and growth of data exclusively in Government sector should be an added advantage for the Government to increase the service delivery to the public. Big Data Analytics (BDA) is one of the most advanced technologies to analyse data owned by the Government to explore other fields, or new opportunities that can bring benefits to the Government. Although BDA concept has been implemented by many parties, there exists a number of misconceptions related to the concept from the aspect of understanding and implementation of the project. National Hydraulic Research Institute of Malaysia (NAHRIM) as one of the four agencies that have been implemented Malaysia's BDA Proof-of-Concept (POC) initiative is no exception to these misconceptions. In this paper, we will discuss the misunderstandings and challenges faced throughout our BDA project, in encouraging and increasing the awareness of the implementation of BDA in Government sector.

1 INTRODUCTION

Big Data Analytics (BDA) has been escalating in various sectors as it increases the value of data in organisations for different purposes. The awareness and understanding of BDA among top management has been familiarised to ensure how data can be analysed, improved and enriched to become a new key economic factor that can alleviate an organisation's performance.

Nowadays, providing and sharing data either internally or externally is less challenging compared to analysing the data. In BDA, analysing data requires knowledge, insight, and wisdom from Subject Matter Experts (SME), specifically in the chosen domain. To embrace the understanding requirement from SME towards BDA's targeted output and outcome is another challenging process.

As mentioned by Phillip Russom in his TDWI Best Practices Report "Big Data Analytics" (2011), the hottest new practices in Business Intelligence (BI) today is BDA. BDA can be achieved by putting massive amounts of detailed information and advanced analytics together. BDA is not just the upgrade and expansion to legacy systems and algorithms, it requires a new set of tools to determine relevant data and to convert this data into useful knowledge (Bi and Cochran, 2014). Companies need to reconsider their methods at the system level strategically, operationally and culturally for data management, and then select the right data, and make right decisions based on it (Troester, 2012).

Business owners typically use BDA to explore other fields or new opportunities that can bring benefits to them. However, this goal is difficult to achieve if the wrong understandings and techniques are used to perform BDA. The misconceptions, technically or theoretically, will give a major impact on the overall analysis process, result and outcome of the domain case.

This paper presents the National Hydraulic Research Institute of Malaysia (NAHRIM)'s participation in Malaysia's BDA Proof-of-Concept (POC) initiatives in visualising 90 years of projected rainfall corresponding runoff after-effects based on river basin in Malaysia. We will discuss the misunderstandings and challenges faced through our experience in BDA project with objective to encourage and increase the awareness of the

Abdullah, M., Ibrahim, M. and Zulkifli, H.
Resolving the Misconceptions on Big Data Analytics Implementation through Government Research Institute in Malaysia.
DOI: 10.5220/0006293902610266
In *Proceedings of the 2nd International Conference on Internet of Things, Big Data and Security (IoTBDS 2017)*, pages 261-266
ISBN: 978-989-758-245-5

importance of BDA implementation in organisations, especially in Government sector.

2 RELATED WORK

Big data creates a radical shift in how we think about research, thus reframing key questions about the constitution of knowledge, the processes of research, how we should engage with information and the nature and the categorisation of reality (Boyd and Crawford, 2012). Kaisler et al., (2013) define "Big Data" as the amount of data just beyond technologies capability to store, manage and process efficiently which the limitations are only discovered by a robust analysis of the data itself, explicit processing needs, and the capabilities of the tools used to analyse it.

There are three aspects to characterized Big Data; the data must be numerous; the data cannot be categorized into regular relational databases; and the data are generated, captured and process very quickly (Khan et al., 2014). Data can be obtained from any source, in various forms by various criteria. The source of big data is basically categorized into two categories, namely data from the physical world and data from human society (Jin et al., 2015). Data from the physical world is usually obtained through sensors, scientific experiments and observations, such as biological data, neural data and remote sensing data, while data from human society is acquired from sources or domain as social networks, Internet, health, finance, economics and transportation.

As for NAHRIM, a government research institution (RI) focusing on research and development (R&D) for water and environment, holds numerous water related and climate change data for Malaysia either primary or secondary data, collected through sampling activities, modelling, simulation and other R&D activities. Those data are being used for water and environment planning, supporting decision making and identifying new potential R&D areas that can be diversified into various domain such as data projection analysis, climate change impact, sea level rise projection, hydroclimate and water resources related issues (Zulkifli et al., 2015).

In 2013, the Prime Minister of Malaysia officially announced the Malaysia BDA initiatives. Malaysian Administrative Modernization and Management Planning Unit (MAMPU) have been appointed to lead the BDA projects for public sector which started with five pilot projects as a proof-of-concept (POC) approach. Four agencies have been involved in developing the POC projects and NAHRIM was one of them. NAHRIM's objective for this POC project titled "Projected Hydroclimate Data Analysis & Visualisation for Potential Drought & Flood Events in Peninsular Malaysia" was to develop a BDA system that will be able to assist NAHRIM in visualising and analysing almost 1450 simulation-years of projected hydroclimate data for Peninsular Malaysia based on 3888 grids.

There are two teams involved in this project, which were BDA Technology Team and SME Team, to ensure the success of the project. BDA Technology Team is responsible to provide technology consisting of the hardware, software and customization services to develop the system. Meanwhile, SME Team for this project are the backbone or the brain of the project that provides the solution, methodology and algorithm regarding the domain of the chosen business case. SME Team for the project was a combination of various background of education and experience that come from researchers, engineers and officers of NAHRIM's Water Resources and Climate Change Research Centre and Information Management Division.

Data input for this project were daily projected rainfall data from the year 2010-2099 based on 6km x 6km grid, projected runoff data from year 2010-2099 based on 6km x 6km grid, and projected streamflow data from year 2010-2099 based on selected location in the vicinity of river basin. These data were consumed by the BDA technology to perform analysis as to provide visualisation of drought for Peninsular Malaysia, to visualise rainfall pattern, magnitude and storm centre and its corresponding runoff pattern and magnitude, and to provide searching and linking function to visualise user-defined period of rainfall event to identify storm centres with the corresponding runoff and streamflow data. NAHRIM BDA POC was developed to assist Water Resources and Climate Change Research Centre researchers to monitor and search projected rainfall and runoff data for the years 2010 to 2099 where the process to identify those pattern required longer time to process without BDA technology.

3 MISCONCEPTIONS OF BDA

3.1 The Thought of Completing the Vs

Often when we hear or read about BDA, it will be associated with various terms that starts with the letter V; Volume, Variety, Veracity, Velocity, Value, Variability and so on. Several years ago, when undue attention for Big Data focused only on size, Gartner group proposed the famous “3Vs”; Volume, Velocity, and Variety. Then, IBM pushed for adding a 4th V; Veracity and this has been accepted by most (Jagadish, 2015). Further Vs have also been suggested such as Variability, Validity, Volatility, Visibility, Value, and Visualisation. However, these are met critically as they do not necessarily express qualities of magnitude. (Li et al., 2016).

Some perceived these Vs as a definition and concept of BDA, while others use it as a requirement for conducting the analysis. These Vs represents the form of data that the researchers, academician, stakeholders or business analyst want to analyse and it is not necessarily need to use all the Vs to conduct a BDA. For example, a study that was conducted by Assunção et al., (2015) for the development of BDA in the cloud surrounds only three Vs namely Variety, Velocity and Volume and consider the other Vs deserve a study on their own.

In the case of NAHRIM’s BDA project, the thought of complying all the Vs was the first issue raised and mandatorily be fulfilled. NAHRIM’s assumption was, BDA project will not be successfully developed if our data did not fall in every criterion of the Vs for BDA. But throughout this project, NAHRIM decided to focus on optimising and exploring our 10 billion hydroclimate projected data in structured format. We were proven wrong by the results of the analysis where Volume, Velocity and Visualisation is more than enough to give us the outcomes that NAHRIM required.

However, BDA is not simply a matter of injecting additional scale, variation, speed or noise to research data sets (Abbasi et al., 2016). What may be deemed BDA today may not meet the threshold in the future. Each of them raised its own individual issues and it is described in the table below which covers the first 4 Vs; Volume, Variety, Velocity, and Veracity:

Table 1: Definition and issue on 4 Vs.

The V	Definition (Wamba et al., 2015)	Issue (Li et al., 2016)
Volume	Large volume of data consume huge storage or consist of large number of records.	Raised data storage and massive analysis issue
Variety	Data generated from greater variety of sources and formats, and contain multidimensional data fields.	Complex structures of data calls for more efficient models, structures, indexes and data management strategies and technologies.
Velocity	Frequency of data generation and/or frequency of data delivery.	Require the speed of data and the speed of data generation matching to meet demand.
Veracity	Inherent unpredictability of some data requires analysis of big data to gain reliable prediction.	Rising issues in quality assessment of source data and how to statistically improve the quality of analysis result.

Understanding our own data is the key of BDA implementation. Based on the data, organisation should identify the expected result required and what are the processes involved including data to be used, type of data, technology to be used, the technique to store the data, the method to process the data and how to integrate them and so on to gain insights and depth to solve real problems. Two data sets of the same size may require different data management technologies based on their type, technological advances allow firms to use various types of data, real-time analytics and evidence-based planning is a growing need to meet the unprecedented rate of data creation, and special tools and analytics needed to deal with imprecise and uncertain data (Gandomi and Haider, 2015).

3.2 The Role of IT in BDA

BDA is the latest technology trend and nearly all organisations have interest to apply this technology for their business analysis. But most of the organisations believes that BDA elements comprise of technical tools only. The truth is, business owners should be involved at all events of BDA development to solve problems related to their business and field. Most of them failed to notice that the main concern for business owners it to provide the sufficient tools and highly trained personnel to work with BDA (Jin et al., 2015).

Business owners used to keep the data by themselves in the past, but for this new era of business strategies and industrial purposes such as the possibility to collect and mine data for desirable information, they should cooperate with the scientific research for the development of their industry (Demchenko et al., 2013). Such data includes market prediction, customer behaviour predictions, social groups activity predictions, and so on. By uncovering business data to the technical consultant, the technical team can assist on proper analysis techniques and technology that can be used to process these data which can save time, cost, and skills. In addition to technical system implementation, significant business, or domain knowledge as well as effective communication skills are needed for the successful completion of such Business Intelligence and Analytics (BI&A) projects (Chen et al., 2012)

IT function is tasked with managing and integrating data as an "enable" of data-driven business processes and decision making (Abbasi et al., 2016). There are numerous number of domain studies that can leverage big data analytics such as atmospheric science, scientific research, government, natural disaster and more (Sagiroglu and Sinanc, 2013). In NAHRIM's case, we are implementing BDA concept through development of an application that can centralize the information and analysis on a web application. NAHRIM took advantages of the current BDA technology by outsourcing the application development phase to technology provider while NAHRIM focus on providing technical advice and content that consist of data, methodology and algorithm in the analysis phase of the project. Our BDA project will not be successful without collective collaboration among researchers and engineers in providing the data and it is not subject to technological problems alone.

3.3 The Dispersion of Data

Unstructured data is one of the data type that revolved around us most rapidly and increasingly. This data type mostly is random and not modelled. Example of unstructured data are audio, video, images, text and human language. Therefore, there is an assumption that only unstructured data is used for BDA. To disabuse this thought, De Mauro et al. (2015) stated that data generated today are increasing in type. Structured data is now joined by unstructured data and semi-structure data. However, the format of semi-structured data does not conform to strict standard spanning a continuum between fully structured data and unstructured data (Gandomi and Haider, 2015).

As a data provider, NAHRIM is no exception into thinking that combination of unstructured and structured data is necessary to be used for BDA. The initial form of data we collect such as rainfall, runoff, and streamflow, are structured data used for hydroclimate projection for the years 2010-2099. These structured data are then analysed through algorithm accelerated by technology provided by technology provider. Through this data processing, we are able to produce data output such as drought visualization by state, month and year, rainfall patterns, magnitude of storms and so on. In this case, NAHRIM do not intent to use unstructured data since NAHRIM would like to focus on optimising the current structured data. This experience can refute the notion that combination of structured and unstructured is a must to perform BDA.

For many organisations, appropriate strategies must be developed to manage such data. Traditionally, data is stored in a highly structured format to maximize its informational contents but because of the current data volumes are driven by these data formats (structured, semi-structured and unstructured), end-to-end processing can be impeded by the translation between structured data in relational systems of database management and unstructured data for analytics (Khan et al., 2014). Unlike the structured data that can be handled repeatedly through a RDBMS, semi-structured data may call for ad hoc and one-time extraction, parsing, processing, indexing, and analytics (Chen et al., 2012). Unstructured data on the way around can be transformed to structured data using Extraction, Classification, Repositories Development and Data Mapping processes (Abdullah & Ahmad, 2015).

A comprehensive research has been made by Yaqoob et al., (2016) regarding of data processing tools including batch processing that can be very efficient where data is collected, stored, processed and results are produced in batches; and stream processing that focus on the velocity of data and help to process data in a very short of time. Example of batch-based processing tools are Hadoop, Skytree Server, Telend Open Studio, Jaspersoft, Dryad, Pentaho, Tableau, and Karmasphere and the stream-based processing tools available are Storm, Splunk, S4, SAP Hana, SQL stream s-Server and Apache Kafka. Selection of big data processing tools is critical as it depends entirely on the needs of users and the type of data that the organisations have. If you apply Extract, Transform and Load (ETL) and data quality processes to big data as you do for a data warehouse, you run the risk of stripping out the very nuggets that make big data a treasure trove for advanced analytics (Russom, 2011).

3.4 The Relevancy of BDA

Most BDA use cluttered data, and not all data is valuable for analysis. Because of this criterion, the process employed to analyse the data obtained are often time-consuming starting from the data collection, data processing and data visualisation. Hence, there is a perception that the end result of BDA is only the data visualisation on the dashboard.

The right thought should be “What are the accurate action that business owners and organisations need to take with the analysed data”. BDA contains a wealth of societal information and can thus be viewed as a network mapped to society (Jin et al., 2015).

Based on NAHRIM experience in applying BDA, we only provide data for analysis. The information that has been generated through this technology hopefully can help the ministries, government departments and agencies such as Ministry of Natural Resources and Environment, Ministry of Energy, Green Technology and Water, Ministry of Agriculture and Agro-based Industry, Department of Public Works, Department of Irrigation and Drainage, State Governments and Private sector to make a careful planning and immediate action in a holistic manner that lead to sustainable development and climate resilience such as water management issues, drought and flood. As mention by Li et al., (2016), geospatial big data has great potential to benefit many societal applications such as climate change, disease surveillance, disaster response, monitoring critical infrastructures and transportation.

On the other hand, BDA can help the people to perceive the present and to predict the future. One of the advanced analytic used nowadays to conduct BDA is analytics continuum by Gartner (Bertram, 2013), which explains the analytic styles from descriptive through to prescriptive. Data or content is being examined to answer the question “What is going to happen?” or more precisely, “What is likely to happen?” This model consists of descriptive analytics, diagnosis analytics, predictive analytics and prescriptive analytics. With the right question and the right data, business owners can obtain more than the analysed data itself. Because of this awareness, several countries have also come up with their initiatives in applying BDA for their national development. Table 2 overview the initiatives taken by international countries on big data based on Jin et al. findings.

Table 2: International Initiatives on Big Data Analytics (Jin et al., 2015).

Country	Event	Initiatives on Big Data
USA	Obama Election Campaign	Find voters and analysed the tendency for voters to vote through real-time data collection and analysis in order to beat Romney and to get re-elected.
	Big Data Research and Development Initiative	Strategic plan that promotes US to continuously lead in high tech field and to protect its natural security through big data research and applications.
British	Big-data plan of £189 million	Aims to push new opportunities for using big data in commercial enterprises and research institutions (on medical, agricultural, commercial, academic research, and other areas)
France	Digital Roadmap (€11.5 million)	Support the development of seven future projects including big data.
Australia	Australian Public Service Big Data	To promote the service reformation of public sectors by making use of big data analysis, developing better public policies and protecting citizen privacy.
Japan	The Integrated ICT Strategy for 2020 Declaration to be World's Most Advanced IT Nation	Plan to develop Japan's new national IT strategy with open public data and big data as its core during 2013-2020.
European Commission	Horizon 2020 (€120 million)	Framework program for research and innovation on big data-related industrial research and application.

4 CONCLUSIONS

Based on NAHRIM experience implementing BDA POC project, our domain business case lead us to focus on leveraging our own projected hydroclimate structured data amounting to 10 billion to assist various entities either Government or Private sector in disaster management in water related disaster such as flood and drought. By understanding the concept of BDA, NAHRIM has proved that BDA is not only about technology, data or problems but the concept requires a full commitment and broad perspectives to understand what we have and what we want to achieve. Today, data are not just grow in the form of quantity but the evolution of data plays an important role that makes data is the new commodity or asset influencing the decision making process. Through NAHRIM's BDA project, it shows and indicate with a correct data, people and technology, BDA concept can be implemented especially in Government sector despite that there are challenges and confusions towards the understanding of BDA concept itself.

ACKNOWLEDGEMENTS

This project was supported by MAMPU, Malaysia Digitial Economy Coporation (MDEC) and MIMOS and we are thankful to our team members from NAHRIM as well who provided expertise that greatly assisted the implementation of this project.

REFERENCES

Abbasi, A., Sarker, S. and Chiang, R.H., 2016. Big data research in information systems: Toward an inclusive research agenda. *Journal of the Association for Information Systems*, *17*(2), p.3.

Abdullah, M.F. and Ahmad, K., 2015. Business intelligence model for unstructured data management. In *Electrical Engineering and Informatics (ICEEI), 2015 International Conference on* (pp. 473-477). IEEE.

Assunção, M.D., Calheiros, R.N., Bianchi, S., Netto, M.A. and Buyya, R., 2015. Big Data computing and clouds: Trends and future directions. *Journal of Parallel and Distributed Computing*, *79*, pp.3-15.

Bertram, I., 2013. Business intelligence: what are you really investing in?, viewed 2 November 2016, <http://istart.com.au/opinion-article/business-intelligence-what-are-you-really-investing-in/>

Bi, Z. and Cochran, D., 2014. Big data analytics with applications. *Journal of Management Analytics*, *1*(4), pp.249-265.

Boyd, D. and Crawford, K., 2012. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, *15*(5), pp.662-679.

Chen, H., Chiang, R.H. and Storey, V.C., 2012. Business Intelligence and Analytics: From Big Data to Big Impact. *MIS quarterly*, *36*(4), pp.1165-1188.

De Mauro, A., Greco, M. and Grimaldi, M., 2015. What is big data? A consensual definition and a review of key research topics. In *AIP Conference Proceedings* (Vol. 1644, No. 1, pp. 97-104).

Demchenko, Y., Grosso, P., De Laat, C. and Membrey, P., 2013. Addressing big data issues in scientific data infrastructure. In *Collaboration Technologies and Systems (CTS), 2013 International Conference on* (pp. 48-55). IEEE.

Gandomi, A. and Haider, M., 2015. Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), pp.137-144.

Jagadish, H.V., 2015. Big data and science: myths and reality. *Big Data Research*, *2*(2), pp.49-52.

Jin, X., Wah, B.W., Cheng, X. and Wang, Y., 2015. Significance and challenges of big data research. *Big Data Research*, *2*(2), pp.59-64.

Kaisler, S., Armour, F., Espinosa, J.A. and Money, W., 2013. Big data: issues and challenges moving forward. In *System Sciences (HICSS), 2013 46th Hawaii International Conference on* (pp. 995-1004). IEEE.

Khan, N., Yaqoob, I., Hashem, I.A.T., Inayat, Z., Mahmoud Ali, W.K., Alam, M., Shiraz, M. and Gani, A., 2014. Big data: survey, technologies, opportunities, and challenges. *The Scientific World Journal*, *2014*.

Li, S., Dragicevic, S., Castro, F.A., Sester, M., Winter, S., Coltekin, A., Pettit, C., Jiang, B., Haworth, J., Stein, A. and Cheng, T., 2016. Geospatial big data handling theory and methods: A review and research challenges. *ISPRS Journal of Photogrammetry and Remote Sensing*, *115*, pp.119-133.

Russom, P., 2011. Big data analytics. *TDWI Best Practices Report, Fourth Quarter*, pp.1-35.

Sagiroglu, S. and Sinanc, D., 2013. Big data: A review. In *Collaboration Technologies and Systems (CTS), 2013 International Conference on* (pp. 42-47). IEEE.

Troester, M., 2012. Big Data Meets Big Data Analytics: Three Key Technologies for Extracting Real-Time Business Value from the Big Data That Threatens to Overwhelm Traditional Computing Architectures. SAS Institute. *SAS Institute Inc. White Paper*.

Wamba, S.F., Akter, S., Edwards, A., Chopin, G. and 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*, pp.234-246.

Yaqoob, I., Hashem, I.A.T., Gani, A., Mokhtar, S., Ahmed, E., Anuar, N.B. and Vasilakos, A.V., 2016. Big data: From beginning to future. *International Journal of Information Management*, *36*(6), pp.1231-1247.

Zulkifli, H., Kadir, R.A. and Nayan, N.M., 2015, November. Initial user requirement analysis for waterbodies data visualization. In *International Visual Informatics Conference* (pp. 89-98). Springer International Publishing.