INTERNATIONAL JOURNAL OF ACADEMIC RESEARCH IN BUSINESS & SOCIAL SCIENCES

Vol 12, Issue 3, (2022) E-ISSN: 2222-6990

Modelling Dark Data Lifecycle Management: A Malaysian Big Data Experience

Ahmad Fuzi Md Ajis[a], Sohaimi Zakaria[b] & Abdul Rahman Ahmad[b]

[a]Faculty of Information Management, Universiti Teknologi MARA, Kampus Segamat, Johor, Malaysia, [b]Faculty of Information Management, Universiti Teknologi MARA, Kampus Puncak Perdana, Shah Alam Malaysia

Email: ahmadfuzi@uitm.edu.my, sohaimiz@uitm.edu.my, arahaman@uitm.edu.my

To Link this Article: http://dx.doi.org/10.6007/IJARBSS/v12-i3/12363 DOI:10.6007/IJARBSS/v12-i3/12363

Published Date: 24 March 2022

Abstract

Qualitative research using 18 case studies were conducted as it allows in-depth investigation and to derive as rich evidence as possible from the selected cases. The data were collected using semi-structured interview on Malaysian Small and Medium Enterprises (SMEs) and the interviews were recorded and transcribed. The transcribed data were analyzed using Grounded Theory Methodology to identify emerging theory on dark data. The purpose of the paper is to investigate the dark data phenomenon towards Small & Medium Enterprise in Malaysia in relation to the scenario of dark data as experienced by SME in Malaysia in relation to its handling consequences towards business entity. The research findings elucidate how Malaysian SMEs dealt with the dark data phenomenon's occurrences which outlined the new model of Dark Data Lifecycle Management. There is a dearth of literature in the area on dealing with dark data, which demonstrates that dark data epistemology is still emerging. Thus, based on the experiences of Malaysian SMEs, a theory was modelled to demystify the dark data lifecycle management.

Keywords: Dark Data Lifecycle Management, Big Data, Small Medium Enterprise, Grounded Theory, Data Abundance, Data Mining, Discoverability And Usability, Malaysia.

Introduction

Exponential advancement of digital devices, applications and connectivity boosts the dissemination of information faster to wider coverage recipients. The leverage of technology makes up creator of information become anonymous and enormous amount of data being spread all day long. The voluminous existence of data and information creates the phenomenon of Big Data with massive volume of data, variety information formats, velocity, and value. This enormous size of big data was analogized as a big chunk of iceberg while the data resides far beneath the stored data line of sight become a mystery to the enterprise (Martin, 2016).

A breached to the data access while some owners of data acknowledge them as their trade secret (once being valued) provide unsafe environment to the survivability of the organization additionally towards its reputations. While some data remains hidden and mystery to the organizational reach which termed as dark data (Berghel, 2007; Northwoods, 2017; Neff, 2018; Corallo, Crespino Vecchio, Lazoi & Marra, 2021), the sheer volume of these mystery data impacts the costs for searching and producing appropriate information and imposes a wasted storage cost in operating budgets (Commvault, 2014; Martin, 2016; Veritas, 2017). Corallo et al (2021) systematically reviewed publications from academic and non-academic institutions upon dark data research but none were indicating SMEs intervention upon dark data occurrences. Presumably, daily routine of handling business transaction data of SMEs, with the domination of business owner over the business activities provide opportunity of dark data accumulation from the data handling procedure occurred during or after transaction. These was identified as the gap to be filled in by this paper.

Therefore, in this paper the researchers intend to uncover the emerging theory of dark data from the perspective of SME in Malaysia using Grounded Theory Methodology. The objectives of the case studies are:

a. to develop a new theory of Dark Data Lifecycle Management based on how SMEs dealt with the dark data, and;
b. to construct the model of Dark Data Lifecycle Management based on the theory.

Emergence of Dark Data: Literature Review

Researcher awareness on the epistemology of dark data was initiated by analyzing the published literatures. Google Scholars, Scopus and Web of Science were utilized to find existing literatures pertaining to the area understudied. The literature searches conducted faced a great challenge whereby publications found on recent 5 years range exhibited limited numbers of literatures. Therefore, ranges of literature searches expanded up to any date of publication which indicate the term "dark data" in the article. The search results appear more convincing whereby a total of 7,020 search hits were found, unfortunately only 56 journal articles and 17 non-academic articles were included to be reviewed after excluding similar literatures indexed by all databases, discarding non-English literatures and availability of full text articles. Review on those literatures exposed major highlights on the current state of the dark data which are (1) volatility definition of dark data; (2) dark data causes; and (3) limited dark data management approach. Volatility of dark data definition due to lack of research on dark data phenomenon only display dark data definition from the perspective of searchability which involved metadata and categorization (Björnmalm, Faria, & Caruso, 2016; Gimpel, 2020; Schembera & Durán, 2020), and usability (Heidorn, 2008; Patil & Siegel, 2009; Brooks et al., 2016; Hawkins et al., 2020). Factors contributed to the occurrences of dark data were found to be limited to the process of ensuring data quality which regards to data accessibility, accuracy, and traceability (Hitachi, 2013; Gartner, 2014; Intel, 2018; Rao, 2018). Although dark data inferred as critical data which exists beyond what is routinely captured and analyzed (Intel, 2018) approach on dark data management were found to be very limited. Only few scholars and industry players suggested dark data can be managed properly by the creation of Data Lake infrastructure using schematic methodology (Trajanov, et al., 2018); four phases of data management including identification, classification, controls, and continuous monitoring (Commvault, 2014).

Uncovering the mystery of dark data relies on how such occurrences defined by the owners of dark data. Hand (2020) described dark data with the analogy of 15 different types of definition even though some definitions are overlapping, or each is a consequence of the other. While these 15 definitions derived from many incidents and research papers described by Hand (2020), many research papers and white papers would rather use Gartner (2014) definition. Yet, established theory on the dark data field is scarce pertaining to what is dark data, how it is being accumulated and how such management of it would benefitting the handlers. Corallo et al (2021) reviewed 22 publications comprises of academic and non-academic publication pertaining to dark data from various fields and perspective, despite, factors of piling up dark data remain ambiguous. Investing on solving dark data issues becomes a tough sell even for large firms (Gimpel, 2020). Besides, there were also 16 white papers on dark data research were reviewed but none were published by small and medium enterprises although dark data were thought to be beneficial (Martin, 2016; Gimpel, 2020; Hand, 2020).

Dark data is information, collected as a function of an organization's normal operations but rarely or never analyzed or used to make intelligent business decisions (Gartner, 2014). Most of it gets buried within a vast and unorganized collection of other data assets. Some refers dark data "data exhaust," because most of the information consider as overlooked information, even though that data has valuable input to the organization and the portions that aren't of value can be a significant drain on resources, including wasted digital storage space (Martin, 2016).

Dark data immersed with information users and creators as they use any mobile storage devices such as tablets, mobile phones and laptops. However, did everybody aware that the piling up of dark data happened in their devices creates risk for the users? An analogy of an iceberg is a good example on how to explain dark data. Approximately 20% of the iceberg would be the visible is regarded as the data that are actively used and visible to the organizations and users. Surprisingly, the bottom part of the iceberg which is the remaining 80% of it could possibly resides with great opportunity for the organizations and users. However, they are hidden and unexposed which usually being kept for reasons such as backup, heritage, and just-in-case the data is needed in the future (Hitachi, 2013; HighQuest Solution, 2016).

Dark data has been defined from various perspective as (Corallo et al., 2021) reviewed 22 publications comprises of academic and non-academic publication pertaining to dark data definition from various fields and perspectives. The systematic literature review conducted by them is to facilitate establishment of dark data definition according to manufacturing industry. It seems that the establishment of the dark data definition falls under the property of searchability (Kambies et al., 2017), accessibility (DiMatteo, 2021), unknown existence (Lugmayr et al., 2017), uncategorized and ignored data (Intel, 2018) which influenced by its formats and led to unused of data (Trajanov, et al., 2018) yet being hoarded though out time.

Hand (2020) published a book describing the definition of dark data by highlighting the dark data (DD) types based on his experience, events happening in many fields and, some were published. Yet, 15 types of dark data which explained by Hand (2020) were still insufficient to reveal the mystery of dark data whereby all these types only define dark data from the

"missing" perspective as it can be described as missing of meaning, missing from being collected or included, missing from awareness of the creators, users, and the recipients of data.

Malaysian Small Medium Enterprises (SME)

Enterprises in Malaysia defined based on economic activities which classified into categories and sectors based on sales turnover or employment numbers (BNM, 2013). Malaysian enterprises are classified into two generic categories which are:

a. Manufacturing

Manufacturing categories refers to businesses who transform physical or chemical materials into new products. Manufacturing is the business with sales turnover not more than RM50 million or full-time workers of not more than 200 workers.

b. Services and Others

Services and others distinguished by the business sectors whereby services include all services including distributive trade; hotels and restaurants; business, professional and ICT services; private education and health; entertainment; financial intermediation; and manufacturing-related services such as research and development (R&D), logistics, warehouse, engineering etc.

Others refers to three economic activities including Primary Agriculture, Construction, and Mining and quarrying.

Those businesses who exceed the definition of SME endorsed by National SME Development Council (NSDC) are defined as Large Firms and not belong to SMEs cluster (SMECorp, 2020).

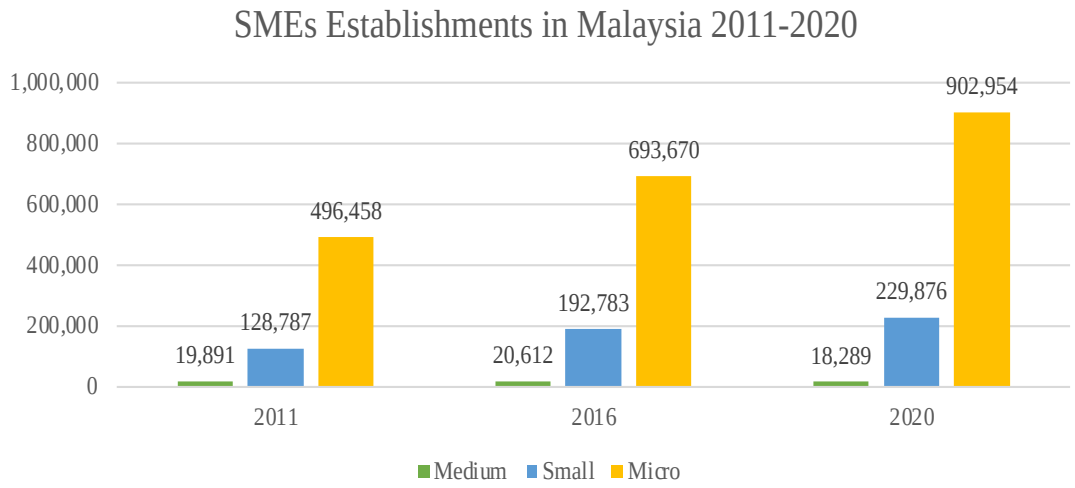

Figure 1. SMEs Establishment in Malaysia 2011-2020

During the past decade, Malaysia SMEs experienced increased in establishment from 2011 with total 645,136 establishment of enterprises to 1,151,119 in 2020. They contributed 38.3% of national Gross Domestic Product (GDP) which consider the second largest contribution which deemed as very much important in the nation building agenda. In 2021, the national budget allocated RM1.9 billion to support the development of SMEs in Malaysia.

Methodology

Qualitative research approach was chosen to initiate the research as there are limited evidence can be found pertaining to the theory related to dark data and lead to the Constructivist or interpretivist philosophical stand to be chose for the study. In this study, SME become the sample for the study because publication of dark data research in journal or white papers publication commonly dominated by large firms or research entities such as Fortune 500 companies in US, or proprietary research firms like Veritas and Ipsos, yet none were found involving SMEs, especially in Malaysia.

Therefore, by using theoretical sampling, 17 Small SMEs was chosen as the cases for the study and expert samples were chosen based on their experience and practices on data handling. This expert sampling method was executed to obtain the expert answers during the data collection. Expert samples for this study were the business owner who have specific expertise on area of the research. They were acknowledged as the expert in the area based on their experience and practices of their enterprise data handling; dominant in the data handling procedure of the company; and involved in analyzing their data and benefitting company's performance.

Data were collected using semi structured interviews guided by interview schedule to keep the interview on track which involved many open-ended questions to open the area of dark data from the experience and perspective of the respondents (Patton, 2014). The data collected were then transcribed and analyzed using Grounded Theory Methodology (GTM). Transcribed interview has gone through the three phases of GTM coding procedure which includes Open Coding, Axial Coding and Selective Coding (Strauss & Corbin, 1998).

Grounded Theory Methodology (GTM)

GTM is the appropriate methods used by researchers to develop a theory for a research field with limited or no theory development (Urquhart, 2013; Charmaz, 2006; Strauss & Corbin, 1998; Flick, 2014; Birks & Mills, 2011). In GTM, Open coding is the phase of opening the area being investigated using conceptualization upon collected data without forcing any preconception to let the theory development grounded by data. The use of gerunds or the "ing" words in assigning concepts and codes really helps in determining processes resides within data (Charmaz, 2006).

Constant Comparative Analysis

Concepts, codes, and categories were generated from constant comparative analysis of shared properties and expansion of dimensions on emerging categories. It is the process of comparing incident to incident, incident to codes, codes to codes, codes to categories and categories to categories which happened continuously during the analysis.

Once the emerging categories and subcategories have been grouped into its hierarchical structure, relationships of each emerging categories were established by employing the Coding Paradigm as guidance (Strauss & Corbin, 1998) during Axial Coding as in Figure 2. Establishing the relationship among emerging categories enables the researchers to find the central categories during Selective Coding which encompasses all the categories emerged. The central category should represent the whole relationships emerged and become the theory emerged from the study.

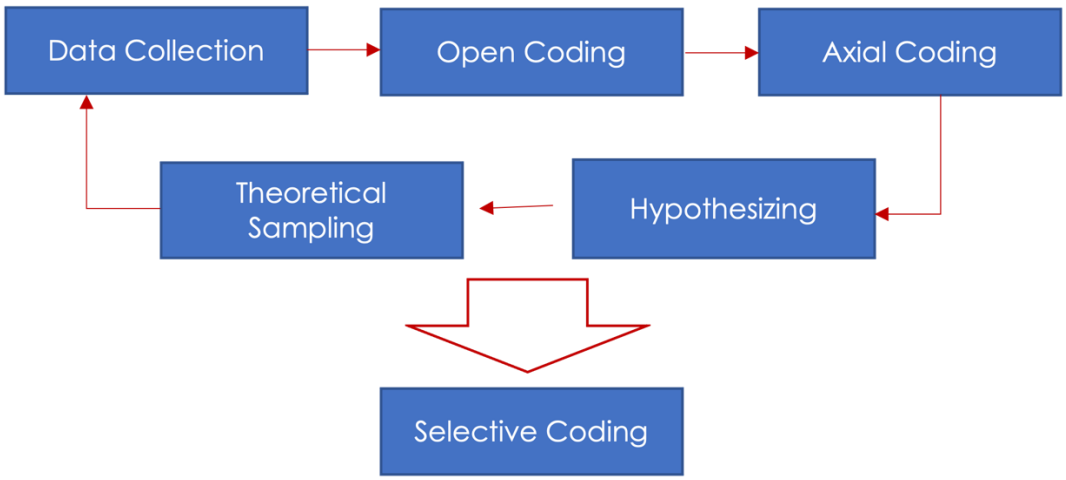

Figure 2: GTM Analysis Process

Findings

Eighteen company owners volunteered to participate in the research, representing a diverse variety of business operations, including manufacturing, wholesale and retail trades, printing and publishing, fitness and health service providers, as well as communication service providers. As stated in Table I, there are three subjects who do business in manufacturing and twelve subjects who offer services while three is from others. Each of the subject are identified in this study as indicated in column ID.

Analysis on Data Lifecycle Management (DLM)

Earth Observation Satellites Committee (2012) assembled 55 ideas and models of data lifecycle management from the standpoint of research, digital data curation, and project data management to form the framework and model of data lifecycle management (DLM). These 55 data lifecycle principles were analyzed and also compared with IBM's six lifecycle stages (IBM, 2013), 7 lifecycle phases of Texas A&M Transportation Institute (Miller et al., 2018) and 6 stages of Big Data DLM (Kumar & Banyal, 2020). The researchers found common similarities of these DLM to be summed up in 4 common stages like Figure 3.

Analysis on the DLM provide significant findings about the study in modelling the dark data lifecycle management (DDLM). It was found that Malaysian SMEs implementing caretaking strategies to deal with the dark data phenomenon which illustrates the similarities of most elements in the DLM.

Emerging Dark Data Lifecycle Management

The standard DLM stages which consist of four processes were found to be extended by the finding of the study. The study reveals that all business owners were employing caretaking strategies which is a strategy to ensure the quality of data by assigning data specialist or caretaker and executing data caretaking or stewarding activities. The caretaking strategies were not only utilized by business owners to manage the residing data in the repository but to suppress the occurrences of dark data phenomenon. This caretaking strategies is modelled using Grounded Theory Methodology and termed as Dark Data Lifecycle Management (DDLM) in this study. DDLM employs 6 stages of lifecycle to manage the dark data as in Figure 4. Although some of the features in the model described below may exist in the present DLM

model, however the DDLM is an attempt to provide comprehensive analytical recommendation of dark data management approach.

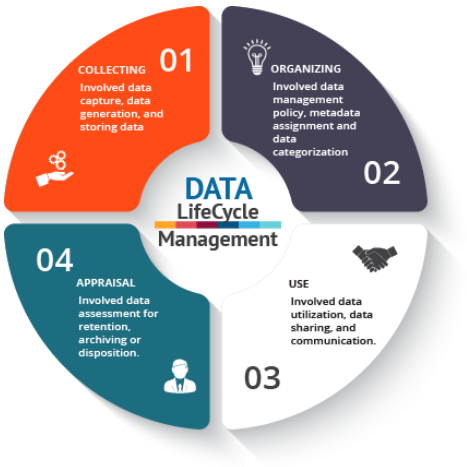

Figure 3: Data Lifecycle Management (DLM)

Emerging Dark Data Lifecycle Management

The standard DLM stages which consist of four processes were found to be extended by the finding of the study. The study reveals that all business owners were employing caretaking strategies which is a strategy to ensure the quality of data by assigning data specialist or caretaker and executing data caretaking or stewarding activities. The caretaking strategies were not only utilized by business owners to manage the residing data in the repository but to suppress the occurrences of dark data phenomenon. This caretaking strategies is modelled using Grounded Theory Methodology and termed as Dark Data Lifecycle Management (DDLM) in this study. DDLM employs 6 stages of lifecycle to manage the dark data as in Figure 4. Although some of the features in the model described below may exist in the present DLM model, however the DDLM is an attempt to provide comprehensive analytical recommendation of dark data management approach.

Figure 4: Dark Data Lifecycle Management (DDLM)

Specializing Caretaking

DDLM initiated by assigning specific data handling responsibility to one or few persons in handling the dark data whereby the knowledge and skill of data handling is crucial for the business activities. Majority business owners only apply in-house data caretaking while some apply hybrid caretaking. The business owners specializing the data caretaking by employing:

a. In-house data caretaking

The use of in-house data caretaking allows the business owners of staff within the enterprise to manage the dark data personally by executing the data caretaking and deciding the procedure of how such data should be approached. Specializing Caretaker implemented by employing in-house data caretaking whereby the responsibility of caretaking the data being assign to specific personnel within the business enterprise. Small SMEs with strong financial resources were found to be assigning data responsibility to staff with close monitoring from the business owner [M-Services IV2; S-Services IV12]

Other than assigning staff as the caretaker, majority subjects dominating data caretaking responsibility of the business organization. Some were believe that it is a must that the business owners owning the data responsibility as they are the person who verify the transaction data, monitoring data analysis and the decision maker [M-Services IV2: M-Manufacturer IV5: S-Services IV15: Mi-Services IV12; IV13; IV14: S-Others IV16;IV17].

b. Outsourcing Data Caretaker

On the other side, external parties also being assigned with data caretaking responsibility as the business owners would like to focus on another side of operation within the company.

The process executed by hiring external party with the purpose of data caretaking responsibility outside the locality of the business organization whereby the business owners hire data responsibility personnel to obtain data driven advice and performance monitoring. Therefore the task of data management from its collection phase until its disposition done with major assistance of hired data managers [M-Services IV2; M-Services IV3].

Collecting

Collecting refers to the activity of procuring data during business activity whereby business owners involved in the process of capturing dark data via multiple platform, generating data from transaction created, gathering data for specific business purposes, and keeping the data within storage facility. The collecting activities practiced by the subjects were structured properly by data procurement and data saving activities due to impact of dark data whereby uncollected data could lead to be missing or lost. Collecting activities involved:

a. Procuring Data

Rich data were created or generated from daily business transaction and the purpose of procuring the data is to capture the data either manually or digitally to record all data involved in business transaction. The recorded data was secured by registering the data in tangible medium like logbooks, digital storage platform or cloud storage which acquired from business activities. Metadata of the recorded data were also described in the register and these data were prioritize to be kept and made available for the purpose of future reference [S-Services IV1: M-Service IV2]. Procuring the data served as the basis of saving the data in the enterprise repository.

b. Saving Data

Saving data refers to the process of ensuring data in any formats to be in possession whereby business owners involved in the activities of gathering and keeping of data during their business activities to enable the data to be referred and used in the future. Multiple storage technology were used to accommodate the data storing process.

Saving the data involved the process of gathering data where existing data were assembled and grouped according to specific reason which to be used again in the future. Keeping the data have troubled some subjects whereby mixture of manual and digital storage facility were used. This issue creates uncertain challenge of data retrieval and data discovery when the data entertained with improper organization procedure by some subjects [S-Services IV1: M-Services IV2: Mi-Services IV6:] Data accessibility is ensured by making it available to be retrieved according to its storage platform features. Some subjects utilized digital storage facility which already equipped with sophisticated retrieval features while subjects with manual storage facility requires more effort in the organizing stage of the DDLM. [S-Services IV7: Mi-Services IV9].

Subjects who already experienced the impact of dark data in the aspect of data quality (accessibility, traceability, retrievability) gathered their data in specific repository to ensure greater data quality. This action allows them not only involved in the activities of putting down data inside the storage but also process of maintaining the storage facility like ensuring data accessibility and protecting them from damages. Backup procedures were designed and initiated by all business owners to prevent any occurrences of damaged data or unreadable

data which enable them to recover those data from their repository [M-Manufacturer IV5: M-Services IV12: Mi-Services IV13].

Organizing

The next step of caretaking activities is to organize the data kept in the enterprise repositories. Organizing refers to the process of cataloguing and classifying data whereby list of data were created and categorized, and separated according to specific requirement. Organization of the data initiated by the subjects after exposed to the dark data except three business owners who already employed DDLM since the day of business establishment [M-Services IV3: S-Services IV4; IV12: Mi-Services IV13]. Organization of data was established after the subjects experienced challenge in accessing and retrieving stored data. It was found that majority micro sized business experienced those unfound data (dark data) due to disorganization which harm their business activities and performance. Therefore, the subjects assigned the stored data with metadata like categories, heading and descriptions as access point to enable the dark data to be retrievable and reduce the occurrences of dark data phenomenon. Organizing process executed by the following sequence:

a. Cataloging the Data

All gathered data were cataloged to create list of the data and sorting them using specific software or platform whereby the data can be utilized and referred for future business activity. During the cataloging process, data were listed according to specific sequence either manually or automatically to overcome the issues of redundancy due to existence of repetitive data [M-Manufacturer IV5: Mi-Services IV11]. Listed data were then given with sufficient description by summarizing or detailing data elements to provide comprehensive information for later searches or reference. The data also described by assigning metadata using specific naming convention to ensure standardization which support effective searching [S-Services IV4: Mi-Services IV10].

b. Classifying Data

Listed data in the catalogue has gone through classification process to assign appropriate headings and categories it helps business owners to identify and recognize the data to be referred and utilized for business activities in the future. Classification of the data was conducted by putting label or heading to the data, grouping and separating data according to specific purposes and approaches. Labelling data were executed based on specific procedure, segregating data based on purpose of business activity and assigning metadata to facilitate searching and analysis process [Mi-Services IV6; IV10].

Use

Data utilization or usage is made enable by the data collection and organization. Accessible and retrievable data allows the subjects to use the data to be applied and to be communicated within enterprise or beyond. The busines owners use the data by the following:

a. Applying data

Applying the data is a situation of business owner take advantage on the data existed and use it to operate the business by consulting to stored dark data to use them for decision making and delivering information or use the data for varied purposes. It

involved analyzing data to identify new innovation ideas [M-Others IV18] and measuring performance [M-Services IV7: Mi-Services IV10: Mi-Others IV16]. Other than analyzing data, business owners consult the data to defend their business from legal claims from the customers [S-Services IV4] and strategizing how customers retargeting should be done for marketing purposes [M-Manufacturer IV5]. Furthermore, the data stored also used as a recovery of inaccessible, hijacked or damaged data [M-Service IV12: Mi-Services IV9; IV11;IV13].

b. Sharing data
Existence of data provide opportunity of business development with the use of connected network especially by leveraging data sharing platform. In this case, data were shared by the subjects to train their staff and notifying updates upon new business strategies and procedures [M-Services IV2: S-Services IV1; IV7: Mi-Services IV6].

Data Mining

Extraction of important information from existing data provide opportunity for the subjects to analyze the data in supporting business decision making and enhance performance. The long-tailed data (dark data) which refers to the old data which rarely being used or ignored due to inactive usage were mined back to facilitate data analytics. These long-tailed data are also including data that is kept for just-in-case purpose as if that it could be needed in the future (Hitachi, 2013; HighQuest Solution, 2016). From the data mining, a lot of knowledge and information were extracted from the data and provide the business owners with opportunities to understand the holistic condition of the business by evaluating pattern of data. At the same time, the subjects used the mined and analyzed dark data to predict the future and strategize the business. This is the attempt made by the business owner to generate new knowledge from data which is not exist in the standard data lifecycle management.

The results from the data analytics assist the subjects to improvise previous mistakes and plan for better business strategy. Unfortunately, the data analytics process were found crucial only to medium and small size business who understand that data is important for the business expansion and sustainability. Micro size business does not concern on data analytics influence on business performance as they operate their business for the sake of daily survival rather than long term business performance, therefore the data mining activity only familiar to small and medium sized businesses. Data mining stages includes the following processes:

a. Extracting long-tail data
Long-tail data is a type of dark data which refers to data that already exist in the repository yet rarely being used or completely being forgotten or ignored. In the data mining stage, the long-tail data was extracted to support data analysis for the business owners to visualize previous and current business performance. The extraction of long-tail data facilitate evaluation of data pattern on profit and capital expenses to enable the subjects in anticipating business' future plan and strategy [S-Services IV1: Mi-Services IV6: Mi-Others IV16].

b. Analyzing Pattern

Based on the extraction of the long-tail data, analysis upon the data generates a lot of information and knowledge that been leveraged by the subjects. There are business owners who able to identify their business problem according to the data pattern and come out with immediate solution [M-Manufacturer IV5: S-Services IV4: Mi-Services IV6: M-Others IV17:IV18]. Results from the data analytics and information extracted assist the business owners to anticipate better decision making for the future operation such as reduction of business cost after learning on the side effect of unrecorded expenses [S-Services IV7: Mi-Services IV11], enhancing customer retention after learning on service quality, [S-Manufacturer IV7: M-Services IV2; IV3: S-Others IV17; IV18] and preventing fabricated data from putting any harm on the business sustainability and profitability [M-Manufacturer IV5: M-Services IV2: S-Services IV1; IV7]. However, the data analytics process were found crucial to all medium and small size business who understand that data is important for the business expansion and sustainability. Some micro size business does not concern on data analytics influence on business performance as they operate their business for the sake of monthly survival rather than long term business performance [Mi-Manufacturer IV14: Mi-Services IV9; IV10: Mi-Others 16] .

Appraisal

Data assessment and appraisal refers to the quality transaction-based data whereby it is determined by the accuracy of transaction process and correctness of the data which free from irrelevant business data and updated to the latest version to be used during business activity. Appraisal processes were conducted based on two major activities which are:

a. Auditing

Auditing is the assessment of the quality transaction-based data whereby it can be determined by the accuracy of transaction process and correctness of the data which free from irrelevant business data and updated to the latest version to be used during business activity. The first process in auditing is verification of data. Verifying data is a process of checking on the accuracy of transaction process executed whereby it influence the correctness of data during business activities. The subjects implement verification of transaction data to control missing data and preventing fabricated transaction by comparing the recorded transaction evidences and updating standard operation procedure.

"We will check the record of received the goods with the money we received. Whether there is something that the shopee forgot to pay. So we can see the details of the correct amount." [S-Services IV1]

"with the SOP earlier we received stock, we know truck driver will send the goods and we will bring one Delivery Order and one invoice, we have to look at the invoice, how much is the value, we have to see how much DO there is then go near the truck, guess how many pallets, then count how many cartons X times Y. just start the forklift, take the forklift, bring in one, arrange this one where first, do you put it on the bottom or put it on the rack "[M-Services IV2]

Afterwards validation of data also in placed to check on the accuracy of data that have been cleanse whereby it determine the pattern of the data to be used later. Validation of data is function to trace missing data by validating data accuracy and very often the process was done manually. Taking the data for granted harm the whole operation and incur additional cost therefore the business owners always do crosschecking the data in hands with the data in their documents of transaction. Although the person incharge of the transaction is the business owners themselves, but mistakes happened at any time and slight error would attract big trouble to the business profitability [M-Services IV2;IV3: Mi-Services IV10].

Next, to ensure verification and validation process is valid business owners prepare some kind of evidence. Therefore, the process to create evidential data of transaction were by establishing standard operating procedure whereby it effective in controlling the occurrences of dark data especially missing data. Evidencing also helps to produce data trails which indirectly prevent data forging and fraudulent activities. Data forging is the events of a party other than the business owners who illegally alter original data and use the data in ordinary transaction which happened to the business owners. These fabricated data harm the business in the aspects of profitability and sustainability which happened to majority of the business owners [S-Manufacturer IV8: M-Services IV2;IV3: S-Services IV1;IV4;IV7: S-Others IV16]

b. Cleansing Data

Cleansing the data is a process of making sure the data in possession are free from irrelevant or invalid data and having the latest version whereby it could avoid misleading information. Cleansing the data involved conducting the weeding process and updating data. Weeding process is the procedure of eliminating unwanted and irrelevant business data by assessing the value whereby it enables the data to be accurate and free from invalid data. Before executing weeding process, the value of data were assessed to ensure appropriateness of data disposition and in consequences protect the quality of data. The process of weeding only involved the business owners although some of them outsource the caretaking responsibility, however the decision of data retention and disposition decided by the business owners. [S-Manufacturer IV8: S-Services IV12]

Indirectly, weeding process simultaneously updating the data. Maintaining the latest updates of data helps to verify inactive data. Disposition of data not only reduce the spaces used in the repository but also notifying the business owners about the actual event happened to the data updates. Updating data status of a customers who already joined to become the business members provide significant advantage towards busines expansion while other kinds of information provide opportunity of learning upon business' challenges and finding an approach to overcome them [M-Manufacturer IV5: M-Service IV12: Mi-Services IV9; IV11;IV13].

Discussion & Conclusion

Malaysian SMEs dealt with the dark data by employing the Dark Data Lifecycle Management (DDLM) to enhance data accessibility, traceability, usability and accuracy. The common DLM is a cycle consist of the 4 processes of how data being collected, organize, use, and appraisal

which happened to the data however, DLM could not accommodate the emergence of dark data phenomenon. Unfortunately, DLM would be one of the reasons of how dark data kept accumulated in data repositories due to the appraisal process which always identified that such data can be kept for just-in-case. Unfortunately, the assessment of records is often ad hoc in systems with inadequate data keeping management (Barragan, 2020). Therefore a lot of data being kept without clear purpose and piled up the dark data.

Unlike the process of DLM, DDLM constructed with unique lifecycle with the existence of two unique stages which are specializing caretakers and data mining. DDLM started with assigning the data caretaker which also suggested by Schembera & Duran (2020) that dark data should be handled by the scientific data officer, a new professionals who responsible for data management. However, in this study, the business owners seem to be the effective data caretaker as they already exhibit successful encounter with dark data. Business owners' tacit knowledge elucidate sufficient insight on how to prevent the occurrences of dark data phenomenon. Existence of DDLM is not to eliminate the existence of dark data completely but to suppress the occurrences of them and minimizing the risk brought with its existence since it is nearly impossible to eliminate the dark data completely.

The next unique feature in DDLM is the stage of data mining as the process of extracting knowledge and information to evaluate data patterns in order to have a comprehensive understanding of the state of the company. Data mining and analysis were also utilised to anticipate business outcomes, as well as to construct company strategies. Miller et al (2020) considered data mining is part of the data usage process, however, according to the finding of this study, data mining is not merely a subprocess of data utilization or data usage but it is a unique process which involved variety data activity such as data extracting (data searching, data filtering, and data identification), data analysis, and data application (Kantardzic, 2020).

Revealing the DDLM seems to be opening the trade secrets of large firms in handling the dark data and benefitting from dark data existence. This research had discover the method of how dark data can be suppress for its occurrences at the early stage, whereby the Malaysian SMEs were found to be mining the dark data and extracting the value for the benefits of the enterprise as what being implemented by CommVault (2014); Intel (2018) and Texas A&M Transportation Institute (Miller et al., 2020). The data mining stage was not found in standard Data Lifecycle Management but included specifically in DDLM to dealt with dark data existence. Extraction and approaches upon abandoned data known as long tail data provides new insight to the advancement in the field of dark data which outlined the element of data mining and develop the DDLM.

The findings of the study depicts that dark data were managed and utilized for the purpose of enterprise growth, enhancing services & profitability, and reducing unnecessary stored data which incurred cost and decrease management efficiency, Surprisingly, among the 3 level of business size only Small and Medium sized of SMEs aware about the negative impacts of dark data and have taken initiatives to manage them and benefitting from its existence. On the other hands, and only few Micro sized SMEs bother to dealt with the existence of dark data and others are accumulating more of it but far from leveraging the dark data.

Although the occurrences of the dark data looks possible to be managed by the DDLM, yet the true challenge is about assuring the data quality in the aspect of accessibility, traceability and accuracy (Hitachi, 2013). Classification of dark data based on the shades of data reveal the actual context of dark data phenomenon (Gimpel, 2020). DDLM might be the gist of how dark data should be encountered by SMEs however, the factors that support the effectiveness of DDLM can be further investigate to achieve totality in dark data management approach.

References

Ahmed, W., & Ameen, K. (2017). Defining big data and measuring its associated trends in the field of information and library management. Library Hi Tech News. 9, 21-24.

Baragan, S. P. (2020). Appraisal and retention of information in the private sector: a case study.(Doctoral Dissertation). Retrieved from https://eprints.qut.edu.au/199783/1/Salvador_Barragan_Thesis.pdf

Birks, M., & Mills, J. (2014). Grounded theory: A practical guide. 2nd Edition. Los Angeles, CA: Sage.

BNM. (2013). Circular on New Definition of Small and Medium Enterprises (SMEs). Retrieved on 2nd July 2021 from https://www.bnm.gov.my/documents/20124/761700/Appendix1-Circular_on_Definitions+_for_SMEs.pdf

Charmaz, K. (2008). Constructing Grounded Theory. 2nd Edition. Sage Publications: London.
Commvault (2014). 5 Ways to Illuminate your dark data. US: Commvault Systems.

Corallo, A., Crespino A. M., Vecchio, V. D., Lazoi, M., & Marra, M. (2021). Understanding and Defining Dark Data for the Manufacturing Industry. IEEE Transactions on Engineering Management,

Delve. (2010). The Essential Guide to Coding Qualitative Data. Retrieved 2021 June 21st from https://delvetool.com/guide

Dimitrov, W., Siarova, S., & Petkova, L. (2018). Types of dark data and hidden cybersecurity risks. DOI: 10.13140/RG.2.2.31695.43681

Erik, J. M. (2016). Dark Data: Analyzing Unused and Ignored Information. econtentMag.com

Gharehchopogh, F. S., & Khalifelu, Z. A. (2011). Analysis and evaluation of unstructured data: text mining versus natural language processing. 2011 5th International Conference on Application of Information and Communication Technologies (AICT), 1-4.

Gimpel, G. (2020). Dark data: the invisible resource that can drive performance now. Journal of Business Strategy

Gimpel, G., & Alter, A. (2021). Benefit From the Internet of Things Right Now by Accessing Dark Data. IT Professional. 23(2), 45-49.

Glaser, B. G., & Strauss, A. L. (1967). The Discovery of Grounded Theory. Strategies for Qualitative Research. Chicago: Aldine

Glass, R., & Callahan, S. (2014). The Big Data-driven business: how to use big data to win customers, beat competitors, and boost profits. Berlin: Wiley.

Guetat, S., & Dakhli, S. (2015). The Architecture Facet of Information Governance: The Case of Urbanized Information Systems☆. Procedia Computer Science, 64, 1088-1098.

Hand, D. J. (2020). Dark Data: Why What You Don't Know Matters. USA: Princeton University Press.

Hitachi. (2013). Big Data - Shining the light on enterprise dark data (EDD). Retrieved April 15, 2019 from https://www.hitachivantara.com/en-us/resources.html

Highquest Solution. (2016). Dark Data Making your Organisation data-enabled? Retrieved on April 4th, 2019 from https://doczz.net/doc/8987800/white-paper-dark-data-making-your-organisation-data

IBM. (2013). The fundamentals of data lifecycle management in the era of big data : How data lifecycle management complements a big data strategy. Retrieved on June 25, 2021 from http://hosteddocs.ittoolbox.com/TheFundimentals.PDF

Intel. (2018). Datumize and Intel transform dark data into operational insight for manufacturing and logistics. Accessed: May 31, 2019. [Online]. Available: https:// www.intel.sg/ content/ dam/ www/ public/ us/ en/documents/solution-briefs/datumize-dark-data-in-manufacturing- and- logistics- solution- brief.pdf

Kantardzic, M. (2020). Data Mining: Concepts, Models, Methods, and Algorithms. Wiley: New Jersey.

Kumar, K., & Banyal R. K. (2020). Data Life Cycle Management in Big Data Analytics. Procedia Computer Science. 173, 364–371.

Mayer-Sch¨onberger, V., & Cukier, K. (2013). Big data: a revolution that will transform how we live, work, and think. Houghton Mifflin Harcourt.

Miller, K., Miller, M., Moran, M., & Dai B. (2018). Data Management Life Cycle: Final Report. Texas, Texas A&M Transportation Institute.

Schembera, B., & Duran, J. M. (2020). Dark Data as the New Challenge for Big Data Science and the Introduction of the Scientific Data Officer. Philosophy & Technology. 33, 93–115

Strauss, A., & Corbin, J. (1998). Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory. Thousand Oaks, CA: Sage

Urquhart, C. (2013). Grounded Theory for Qualitative Research: A Practical Guide. Sage Publications: London.