DEMYSTIFYING DARK DATA CHARACTERISTICS IN SMALL AND MEDIUM ENTERPRISES: A MALAYSIAN EXPERIENCE

AHMAD FUZI MD AJIS[1], SOHAIMI ZAKARIA[1], ABDUL RAHMAN AHMAD[1]

[1]Faculty of Information Management, Universiti Teknologi MARA, Malaysia.

Email: [1]ahmadfuzi@uitm.edu.my

ABSTRACT

The paper reports a study on the dark data phenomenon experienced by Small & Medium Enterprise in Malaysia in relation to characteristics of dark data from SME perspectives. Qualitative research was conducted upon 13 cases which derived from the Inductive Grounded & Emergent Theory Sampling, implemented to identify appropriate samples for the study. Data was collected using semi-structured interviews which were recorded, transcribed, and analyzed using the Grounded Theory Methodology. The research findings highlighted the characteristics and types of dark data typically resides in SME repositories. The study further shows that the field is lacking in literature on dealing with dark data which depicts dark data epistemology are still evolving. Thus, based on Malaysian SMEs' experience a theory was suggested to demystify dark data management.

Keywords: *Dark Data Types, Dark Data Characteristics, Dark Data Management, Grounded Theory, Malaysia.*

1. INTRODUCTION

The phenomenon of Big Data is characterised differently by researchers depending on the subject of knowledge. It is frequently connected with the collecting and analysis of huge or enormous datasets within the owners' repositories [1], and this ambiguous connotation was used to describe Big Data in the absence of a clear definition. The use of an integrated network environment [19], also known as the Internet of Things (IOT), the existence of a social media environment [9], and the habit of generating, analysing, and sharing data all contributed to the emergence of Big Data. The involvement of social media platforms in this phenomenon, whereby social networking were connected not only for personal purposes but also for businesses to reach out to the market and government to ensure its presence in citizens awareness, while many IT specialists connecting multiple devices to be controlled by sensors and portable devices to accommodate the industrial revolution necessities, also contributed to this phenomenon.

The analysis and use of actual data is the primary focus of studies in big data across disciplines [36]. Social networking, business, and governments are typically, but not exclusively, interested in understanding the implications of big data on social behaviour, business practises, and the government's presence in the lives of citizens [21],[31].

Presently, the literature has only given a few examples of the vast world of big data, which is primarily concerned with physical data, data that is readily available to the users [37]. When it comes to physical data, dark data is generally anything kept on a server that's accessible for use but that arguably no one knows about. Yet this description of dark data lacks comprehensiveness especially when it required more investigations to be developed. Data exists at any information users and creators as they use any mobile storage devices such as tablets, mobile phones and laptops. However, unawareness on 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 [32]. 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. Although they are hidden and unexposed they might be usually being kept for reasons such as backup, heritage, and just-in-case situation where the data may be needed in the future. Yet such evidence of these practices is yet to be discovered and such theory is lacking.

Therefore, the researchers attempted to employ the Grounded Theory Methodology to discover a developing theory about dark data. As part of the case study's goals, the researcher look at how SMEs dealt with the management of dark data, as well as variables that support the management approaches.

2. LITERATURE REVIEW

Massive data collections from Big Data phenomenon immersed with multiple data format and structures. Mixtures of unstructured and structured data created information retrieval challenge especially those value and context immersed deep in the unstructured data. Computers can quickly and readily identify and recognise structured data since it has a well-defined structure with a data model [16],[22]. In MS Excel, for example, you could have data from a purchase transaction organised into fields and columns. Structured data is searchable and may be arranged using data types. While structured data has the qualities of well-organized data, unstructured data has the characteristics of unstructured information or meaning which is derived not from its structure but from its values and context. For example, context and values from videorecording or images, can only be extracted if one viewing it and describe them in an understandable content. Organizational data exists in the form of both organised and unstructured data, as well.

There is a difficulty to unstructured data analysis because of the content and data value that is buried inside structured and unstructured data requires human interpretation [9]. Images, movies, and audio in their context provide significant information that can't be accessed by computer keyword analysis in unstructured data. They also encourage the creation of redundant, outdated, and trivial (ROT) information. The inability to access data puts the accuracy of data analysis at risk, which may lead to poor decision-making based on incorrect data.

Few researchers and industry players have proposed many initiative to extract with the hidden values immersed in unstructured content [8],[25],[26] while some developed framework of assessing the images data (unstructured content) which massively resides in organizational repositories and increase the cost of storage yet left from being used [29]. The occurrences of these hidden, inaccessible, and abandoned valuable data is termed as the Dark Data, the data assets that companies gather, analyse, and retain on a regular basis, but fail to utilise for other reasons in most cases [15].

Dark data, according to Dimitrov et al. [12], poses a danger to a company's financial health. Existence of the dark data in the corporate repository may cause issues with data management regulations and legislation compliance. In Malaysia, for example, the Personal Data Protection Act (PDPA) prohibits the storage of personal data for more than seven years unless there are compelling reasons to do so, and those who provide flimsy justifications risk being sanctioned. Unlike active data, which is presently being utilised, dark data sits deep inside an organization's systems and is seldom used or even known about, which leaves data owners vulnerable to cybersecurity threats and risks such as personal data breaches and stolen data, putting the company at danger. Whenever neglected dark data leaks or falls into malevolent actors, the intellectual property rights are jeopardised, company trade secrets are placed in jeopardy, and business intelligence is decreased. Businesses' reputations would be damaged and revenue would be harmed if their data security was compromised, and data would be taken. To add insult to injury, if these companies are unwilling to engage in mining their dark data, they will lose out on opportunities for development including efficiency increases and customer behaviour analyses to improve services and profitability as well as prevent liability. Otherwise, a company would be exposed to open-ended dark data danger as long as it ignores the many possibilities that dark data presents.

Dark Data, a mysterious subset of recorded data, may be found in the gaps between sheets of organised information. It is very uncommon for academics and researchers to produce a wide range of Dark Data definitions that take into account different points of view. However, the concept of "dark data" has not yet been universally agreed upon. Dark data is information that is gathered as part of an organization's regular operations but is seldom or never evaluated or utilised to make informed business choices, according to the few academics who have focused on the topic. Most of it is lost in a jumble of other data assets that is both large and disorganised. Dark data has been referred to as "data exhaust" by some since the majority of the information is seen as unimportant, despite the fact that it contains vital information for the company [23]. In addition, parts of data that aren't useful may be a major drain on resources, including wasted digital storage space [32]. However, the phrase "dark data" has grown ambiguous due to a dearth of theory to explain it.

Dark data should not be defined only by the paradigms of its acquisition or discovery, even if these paradigms become inseparable from dark data. However, even though Corallo, Crespino, Lazoi & Marra [9] proposed that the key concept of definition according to manufacturing industry perspectives define dark data from its description (catalogue), capturing activity, advancement in analysis tools, sources and data formats, but definition of dark data was still inadequately covered. Therefore, the researcher suggested dark data definition may be contested from the following perspective:

a. treatment and existence;
b. dark data shades.

There were also multiple research were done to identify the existence of dark data. Veritas established databerg in 2015 to research dark data. As seen in Databerg, big data is divided into three layers: business vital data, ROT data and dark data. The survey included 1,475 individuals from 14 countries across Europe, the Middle East, and Africa (EMEA). The study concluded that 54% of EMEA firms' data was dark, 32% was redundant, outdated or insignificant, and just 14% was vital business data. Another study was done in 2017, whereby Veritas and DLT launched a new study on dark data using government data [43]. As government data grew, difficulties with dark data arose when the data's information and values were unclear. So 203 public servants were polled to solve the government's dark data issues. The study indicated that more than 60% of government data was expected to remain dark, mostly due to storage capacity difficulties, since data rose by 40% annually while storage capacity only expanded by 9%. Other than that, the Splunk had an investigation found in 2019 that the lack of resources and time to prepare data for utilisation resulted in dark data. The data is hidden because its owners either don't know it exists or don't know how to utilise it. With over 50% of their organization's data going dark, data recovery is difficult. As firms struggle with dark data, over 30% turn to experts for help, yet 25% find their help inadequate.

Even though few publications addressed on dark data's properties and characteristics, the definition gap based on solid theory is still unclear and unexplained by real-world phenomena. There were a large number of academic papers that addressed dark data, as well as research groups that used dark data in their work. But from the non-academic viewpoint, only big companies have released their white papers and research articles on the presence of dark data, its administration, and the ways in which it may be used [9]. Dark data occurrences in the area of big data triggered enthusiastic discovery of the data's influence and present status on companies. Numerous worldwide research firms conducted dark data research in order to unravel the enigma surrounding dark data and the benefits it offered to the global big data community. As a result, undertaking dark data research from the perspective of Malaysia's small and medium-sized enterprises would fill the void.

2.1 Malaysian Small Medium Enterprises (SME)

Malaysian businesses are categorised according to economic activity, such as sales turnover or the number of employees [5]. Malaysian businesses may be divided into two broad categories: manufacturing and services.

a. Manufacturing
Manufacturing companies are those that take raw materials and turn them into finished goods. Manufacturers are companies whose sales turnover is less than RM50 million or that employ less than 200 people on a full-time basis.

b. Services & Others
Businesses distinguished by types of business activities such as distribution, lodging and food service. Other types of services include private education and health, entertainment, financial intermediation, and services related to manufacturing, such as R&D, logistics, warehouse management, and

engineering. Primary Agriculture, Construction, and Mining & Quarrying are examples of the "others" category.

Services are the largest population of SMEs in Malaysia followed by manufacturing and other services. These business entities is of the largest contributor to Malaysia Gross Domestic Products (GDP) in 2020 with more than 38% contributions. Therefore, Manufacturing and Services sectors were chosen for the study as these two sectors are the biggest contributors to the Malaysia GDP by SMEs, and the largest community of enterprises in Malaysia.

In a nutshell, the literature review indicates the hidden values reside in the massive data in unstructured data, which creates the occasion of hidden, abandoned, and unused potential data, known as "dark data," that consumes storage spaces of organisational repositories [9], [19]. Scholars predicted that its existence would be detrimental to business growth and survival [11]. On the other hand, the conclusive definition of dark data is yet to be standardized, and a variety of dark data definitions were found, representing different comprehensions of dark data analogy [9],[11]. Due to the inability to identify the existence of dark data, the tentative definition of dark data causes significant confusion among data owners regarding how to prepare for dealing with such data. Moreover, as scarce publication was found to entail the benefits of dark data in real practice, the danger of dark data existence seems to be emasculating the dark data's untapped potential rather than revealing the leverage it brings [23]. Although few research companies revealed that dark data exists and more than half were identified as such from the data holdings [39], the identification of dark data was still rebounded on the principles of unused stored data assets. In fact, dark data conception is beyond data hoarding. Consequently, while benefitting from dark data remains the trade secret of big firms, this phenomenon has developed a huge gap in dark data utilization, especially for small companies and public sectors.

3. METHODOLOGY

Qualitative approach were taken to investigate the field of dark data and discover the area of research inductively. Sample of population were initiated using purposive sampling known as Inductive Grounded & Theoretical Sampling which accommodate the theory development using grounded theory. This sampling method is also known as theoretical sampling, a method of which sampling are made to densify emerging categories or theories at hand by deciding what data to be collected next, and where to find them. Experts were selected for the study based on their knowledge of data management methods and their prior experiences. Expert samples were used to get expert opinions throughout the data gathering process. which comprises of company owners who considered as expert samples because of their in-depth knowledge in the study's subject area. In the business's data processing method, they had a prominent position, and they were engaged in the analysis of their data and benefited from the success of the company. Therefore, 14 business owners were selected using theoretical sampling to discover emerging theory on dark data management.

Semi structured interviews were employed to gather data and guided by the interview schedule to keep the interviews consistent. The interview questions covered issues on management of business data such as marketing data, staff data, customers data, financial data and operation data. The data collected were recorded using audio recording, and transcribed using Intelligent transcription whereby pauses, filler words, and redundant repetitions were omitted. The recordings made during the orientation stage were not immediately transcribed but were briefly recorded as demographic data about the interviewees. Neither verbatim nor edited transcription were used, as the former transcribes every single word in addition to the filler words that are not needed during the analysis process while edited transcription is usually used to publish the transcript to a specific audience [10]. Some information may be summarised which is inappropriate to be used as some important information may be left out [10]. Transcribed data were analyzed using three stages of Grounded Theory Methodology including Open Coding, Axial Coding and Selective Coding [38].

3.1 Grounded Theory Methodology

The coding process used in grounded theory analysis is divided into three phases, each of which has a distinct emphasis and goal. There are three phases of coding: initial, intermediate, and advanced, which group together all the various names of coding process serving the same function as in table 1, which classified these stages as initial, intermediate, and advanced.

In a nutshell, Initial Coding or Open Coding is the process of slicing up collected data into smaller pieces to identify emerging categories, Intermediate Coding or Axial Coding is the process of establishing relationships between categories and subcategories, and Advanced Coding is the process of developing the theory or central categories that encompass the entire related phenomenon or emerging categories which also known as Selective Coding or Theoretical Coding [40]. Figure 1 exhibit the process of GTM analysis.

The coding procedures are a major task in the process of densifying emergent categories, since the process of forcing ideas or concepts into categories disrupts the theory based in facts, a phenomenon referred to as Theoretical Sensitivity [7],[40]. The researchers' prior knowledge influenced the occurrences of data forgery rather than allowing the data to truly explain what occurred. For example, academic researchers' experience in the field of data management may create an assumption about how data is managed, while actual data management by street practitioners may tell a different story.

Table 1 Different GTM coding Procedure (Urquhart, 2014)

Scholars	Initial Coding	Intermediate Coding	Advanced Coding
Glaser and Strauss (1967)	comparing incidents to each category	integrating categories and properties	writing the theory
Glaser (1978, 1992)	Open coding	Selective coding	Theoretical coding
Strauss (1987); Strauss & Corbin (1990, 1998)	Open coding	Axial Coding	Selective Coding
Charmaz (2006)	Initial Coding	Focused Coding with emphasize on Axial Coding (Coding Paradigm)	Theoretical Coding
Corbin and Strauss (2008)	Open coding	Axial coding	Theoretical coding

During the coding procedure, Constant Comparative Method (CCM) was employed whereby it is the process of comparing concepts that have common characteristics in order to classify them into different categories that provide analytic distinctions [20].

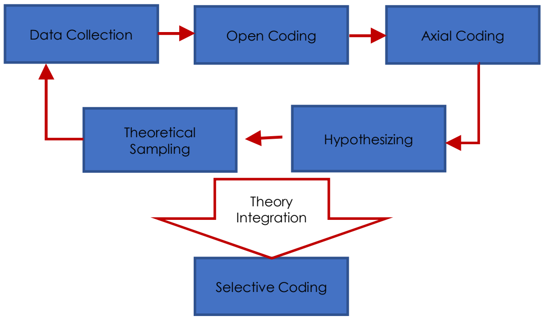

Figure 1: GTM Analysis Process

During the subsequent data analysis, concepts with disparate characteristics were separated and expected to be put along with developing categories. Additionally, CCM aides in determining the dimension of each group.

4. RESPONSE AND FINDINGS

Those who agreed to participate in the study includes 13 business owners from a wide range of business activities, including manufacturing, wholesale and retail trades, printing and publishing, fitness service providers as well as health service providers, as well as service providers of communication. There are three people who venture business in manufacturing, and ten people who provide services, as shown in Table 2.

Analysis using GTM discovered interesting findings for the research question of the study. Firstly, the discovered findings clarify the types of dark data based on the characteristics of the data being studied. Clarification of the dark data types provide useful insight on how business owners dealt with the phenomenon of dark data.

4.1 Phenomenon of Dark Data

The study findings clarified the types of dark data to be classified based on the shades of the dark data. The researchers proposed to classify the phenomenon of the dark data based on the characteristics of the data which in this study coined as Black Data and Grey Data.

Table 2. Responded Malaysian SMEs
N=13

Sector	Size	Business Activities	n	%
Manufacturing	Medium	Printing & publishing [Interview 5]	1	7.7
	Small	Food & Beverages [Interview 8]	1	7.7
	Micro	Food & Beverages [Interview 4]	1	7.7
Services	Medium	Wholesaling & retail trades, and communication service [Interview 2, 3, 12]	3	23.1
	Small	Wholesaling & retail trades, health services [Interview 1, 6, 7]	3	23.1
	Micro	Food & Beverages, Fitness Services. [Interview 9, 10, 11, 13]	4	30.7

Further, each dark data types were described based on its characteristics.

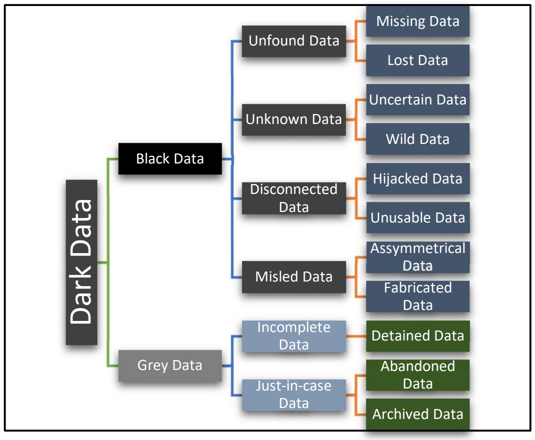

Figure 2. Dark Data Phenomenon

4.1.1 Dark data: black data

Dealing with daily business activities exposed the business owners with the data of which beyond the control of the business owners in terms of possession, searchability, and awareness whereby the usage of data is prohibited, misleading or merely impossible to be utilized. This is a completely dark data situation termed as Black Data. Black data comprises of the following characteristics:

a. Unfound Data

Missing and losing of valuable data are the most experienced events by Malaysia's SMEs and commonly caused by unrecorded data or human error. Unrecorded data happened to majority of the business owners at the early establishment of the business whereby valuable data were not recorded and resulted data missing. Infrastructure failure also contributes to the causal factors of data lost. Missing data provide the opportunity to be recovered as certain data might still resided within repository however, lost data is no longer available and no recovery action can be taken.

"His problem is that we use the product and then we don't pay, and unfortunately we don't record. When we do the stock count something must be lost." [Interview 2]

"Once, the accountant brought back the file. He wanted to create an account for the company. Then, for example, 10 that he borrowed he only returned 7, 3 were missing, then he didn't notice." [Interview 3]

b. Unknown Data

During the operation of the business, a lot of unknown data were remain undiscovered which increase the impact of dark data upon the business performance and operation. The unknown data was derived from the situation of undiscovered data which is not/yet controlled or yet in possession; or in possession but not informed of its existence and usage; or beyond the knowledge of the business owners even though sources of the data is identifiable.

Unknown data were accumulated from the uncertain data which refers to data with inconsistency context or uncertain meaning whereby the value and the influence of the data is unrecognizable and unable to be predicted. Example of uncertain data were marketing data impact on sales, awareness of certain knowledge or data, and inconsistent data output of similar operation.

"Today the ads don't give a lot of impact. There are a lot of leads, but they don't buy, It means I mistakenly target the people.

Sometimes, in one day up to 100 people use WhatsApp, but only 1 person buys it, so we have to change the ads." [Interview 7]

"The bread has to be monitored, although we bake using the same duration, the output is not the same all the time. Because the flour that comes, the raw material that comes is from a factory that has old flour and new flour, so it has gluten, it's not the same. That's it. It's all a secret, we have to be good at looking at it, but it's not far away. The difference, maybe in a minute or two." [Interview 8]

Furthermore, the unknown data also a result of the wild data accumulation, data that presently is not in possession of business owners whereby it is not yet being collected, controlled, gathered, organized or produced but later obtained in variety ways to leverage the benefit from the data. Commonly the wild data were captured for marketing purpose which derived from the untapped community to be targeted market.

" My first marketing funnel is to share first, we use pull marketing that is not profitable. If you ever see an ad that doesn't ask you to buy but they share tips. People nowadays don't like to see us doing hard selling. they love us to share input or knowledge. if you've ever seen a chef, he'll tell his own recipe first, then he sells, he'll attract followers. That's why his sales skyrocketed. 10 minutes and It's all over. " [Interview 7]

"When there's a member join through walk-in, we will ask for his friend contacts, we ask for 3-4 telephone numbers and using that number, we blast text through WhatsApp." [Interview 10]

c. Disconnected Data

At any time the data being collected or being reuse, the occurrences of disconnected data always happened whereby such data could not be used although the business owners were having it, which affected by its relevancy, accessibility, and fragility. The data that were hope to be use in the future suddenly disconnected from its purpose.

Accumulation of data provides competitive advantage to business enterprise and at certain time, the data collected threatened by the safety security. Disconnected data is influenced by the security threats whereby events of Hijacked Data happened. Hijacked data is a situation of data security being breached and the data become inaccessible due to the reason of blocked access, or stolen whereby it may lead to data lost. However, disaster recovery plan prepared in the early establishment of the business would prevent such disaster to happened.

"If my data stolen, my customer will be influenced to change their service provider. "[Interview 11]

"If let's say we cannot access the system, we will access it remotely, we can use the remote desktop because the remote desktop has another user account, if we still unable to access it means it already been hijacked." [Interview 12]

Moreover, Disconnected data contributed by the concept of unusable data whereby data are inaccessible, damaged or corrupted, or not relevant or unable to be used by business owners either by manual utilization or by computer processing.

"We've experienced corrupted data very often, commonly the design data which always corrupted so we just delete it and create a new one. " [Interview 5]

"If the data is in the google sheet, there is no problem anymore, for pictures data if you change the phone, the old pictures will be corrupted or missing." [Interview 6]

d. Misled Data

During business transaction, misleading information could harm the business profitability and disrupt the performance of the business. Incorrect data provided by any parties could disturb the business activities and create additional liabilities to the enterprise.

Misled data derived from the asymmetrical data whereby contextual meaning of data were differently interpreted by the recipient and commonly caused by communication error.

"I used to say that I misunderstood when I shared info via whatsapp, why don't I understand this, why is this different" [Interview 9]

"It's normal when we use WhatsApp, it's normal for him to misunderstand, but we have to correct it." [Interview 11]

Another misleading data is fabricated data which are falsified data created by customers or staff whereby its existence provide harm to the business profitability. The events of fabricated data experienced by majority of business owners and strategies of dealing with this type of dark data were initiated to prevent its repetition in the future.

"He said the money was transferred. After that, we checked the receipt to see if he really transferred but we try to verify the transaction he didn't respond. Maybe he already knew that we knew he was cheating." [Interview 1]

"Let's say the price is RM100, so this staff will charge the customer RM200. the customer paid RM100 in cash, so the staff will take another RM100 without nobody noticing it but she dont realize that every actions in the premise was recorded. " [Interview 4]

4.1.2 Dark data: grey data

The black data is the dark data with the darkest shades whereby it prohibits accessibility, and utilization. Another shades of dark data is termed as Grey Data. Grey data is the opposite characteristics of black data in terms its accessibility and utilization. The data is somewhat lacking of its searchability, inactively being used and kept for certain period although the data is available and its existence is in the know of the owner. Grey data is comprises of incomplete data and just-in-case data.

a. Incomplete Data
Incomplete flow of business transactions or activities generate incomplete data that might influence inaccurate decision making. In this sense, the incomplete data is accessible and able to be used for business decision but its incomplete content or context may direct the information interpretation to be incorrect. Incomplete data is derived from detained data whereby the data being hold up due to the delayed process or business activity whereby it disturb the accuracy of data analysis.

"Actually, it was money from 3 months ago, it messed up the cash flow a bit there. It messed up our strategy a bit." [Interview 4]

"The reason we rarely do promotions because this thing needs to be set up properly. Because the ePay that I am using, there are few settings that I haven't configured yet, so I can't use it yet." [Interview 11]

b. Just-in-Case Data (Long-Tailed Data)
Accumulation of business data requires the data owners to save the data for future use and the just-in-case reasons become the main factor of why all data being kept. Although these just-in-case data is somewhat being acknowledge to be used as contingency data if any data mishap happened, yet the data usage is rare and only used when needed and the capacity of the storage room could be incapacitated gradually.

Just-in-Case Data is the kept data which commonly being abandoned. Its existence is influence on its format of data whether is support simple data interpretation method or requires complex or lengthy procedural analysis. Example of abandoned data is audio recording which requires the user to play using certain equipment and required active listening skills to comprehend and extract values from the data.

"If we send a message on WhatsApp, we can trace it back, we can look for it back, but if it's a voice, it can't be. Indeed, usually we just abandoned the voice note." [Interview 3]

"I used to rarely focus on the account, that's my weakness. I look at the bank account, if it doesn't fall, it's okay, profit, that's my problem. if you don't drop, it means the company is not lost. From before, I was really lazy to look at the account." [Interview 7]

5.0 DISCUSSION & CONCLUSION

Malaysian SMEs dealt with dark data by employing caretaking strategies to enhance data accessibility, traceability, usability, and accuracy.

The process started by assigning the data caretaker, which was also suggested by Schembera & Duran [37] that dark data should be handled by the scientific data officer, a new professional who is responsible for data management. However, in this study, the business owners seem to be the most effective data caretakers as they already exhibit successful encounters with dark data. Business owners' tacit knowledge of the totality of business operations provides sufficient insight on how to deal with dark data to suppress the occurrences of the dark data phenomenon.

Findings from the research provide a glimpse of the dark data types that reside within the repository of the storage facility. Results of the findings were found to be significantly increase the awareness on dark data identification and support business owners to anticipate on dealing with it. The danger that the dark data brings not only influence financial drawbacks but also jeopardize accurate decision making [12]. As larger dark data accumulation consume larger budget on data storage, the obstruction of data accessibility, interoperability and reuse [37] due to the nature of dark data itself would jeopardize the data accuracy in data analysis and could resulting faulty data driven decision making [39]. As a results, more fiscal resources would be wasted and business sustainability would be difficult. Revelation of dark data characteristics in this study also contribute in building the awareness upon dark data whereby it hold the stand to be the precious assets and trade secrets of local SMEs in surviving their business. Identification of the precious dark data resides in the repository of the enterprises provide insight on dark data existence and stimulate reaction on how to deal with dark data occurrences and benefitting from it. It seems that executing the data management lifecycle would be useful to suppress the occurrence of dark data. Although this research discovered the types of dark data and their characteristics, from Malaysian perspectives, there were publications depicts that mining existing dark data and extracting dark data value is beneficial to the enterprise, as being implemented by Commvault [8] and Intel [26].

While dark data is currently remaining in its epistemological gap as a revelation of its theory and management practises lacking research done by Corallo et al. [9] displays unstandardized characteristics and properties of dark data that were differently used by scholars and created a vague conceptualization of dark data that has become a significant gap in dark data research. Although some large firms share tips and tricks for managing dark data based on their experiences [8],[26],[27],[41],[42], clear guidance for dealing with dark data occurrences remains a trade secret for these large firms and becomes dark information for smaller enterprises. Those large firms' initiatives in dealing with dark data require expensive expenses, which small and medium-sized companies assume will be difficult to deal with more expenses in overcoming their dark data issues. Therefore, to overcome this issue, practises in data management should be investigated to deal with the occurrence of dark data.

Despite, the study demystify the characteristics of dark data, the findings was limited by the theoretical sampling employed during the study. The samples were concentrated on Malaysian SMEs, and only a small number of samples were examined to address the study's research objectives. As a consequence, the generalizability of the theory generated based on the research findings may be examined using a wider statistical survey approach. Furthermore, utilization of interview as approach in the data collection was be limited by the respondents perceived knowledge. Therefore, it is suggested that future research efforts involve site observation, which might lead to the finding of more data for theory development.

REFERENCES

[1] 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.

[2] Banafa, A. (2015). Understanding dark data. Retrieved on Oct. 16, 2020. from https:// www.bbvaopenmind.com/ en/ technology/ digital-world/understanding-dark-data/

[3] Berghel, H. (2007). Hiding data, forensics, and anti-forensics. Communications of the ACM. 50(4), 15-20.

[4] Björnmalm, M., Faria, M., Caruso, F. (2016). Increasing the Impact of Materials in and beyond Bio-Nano Science. Journal of the American Chemical Society. 138(41).

[5] 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

[6] Brooks, C.F., Bryan Heidorn, P., Stahlman, G.R., Chong, S.S. (2016). Working beyond the confines of academic discipline to resolve a real-world problem: A community of scientists discussing long-tail data in the cloud. First Monday. 21(2).

[7] Charmaz, K. (2008). Constructing Grounded Theory. 2nd Edition. Sage Publications: London.

[8] Commvault (2014). 5 Ways to Illuminate your dark data. US: Commvault Systems.

[9] 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.

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

[11] DiMatteo, S. (2021). Cloud Services: Deftly Deal With Dark Data. Chemical Processing. 83(3), 40-48.

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

[13] Erik J. Martin, (2016). Dark Data: Analyzing Unused and Ignored Information. econtentMag.com

[14] Flick, U. (2014) An Introduction to Qualitative Research. 5th Edition, Sage Publications, London.

[15] Gartner (2014). Turning Dark Data into Smart Data: How Email and File Level Analytics Can Lead to Greater Business Value in the Age of Information.

[16] 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.

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

[18] Gimpel, G. (2020b). Bringing dark data into the light: Illuminating existing IoT data lost within your organization. Business Horizons. 63, 519-530.

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

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

[21] 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.

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

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

[24] Hawkins, B.E., Huie, J.R., Almeida, C., Chen, J., Ferguson, A.R. (2020). Data Dissemination: Shortening the Long Tail of Traumatic Brain Injury Dark Data. Journal of Neurotrauma. 37(22), 2414-2413.

[25] Hitachi. (2013). Big Data: Shining the Light On Enterprise Dark Data (EDD). Hitachi Data Systems. Retrieved March 5, 2020 from https://cupdf.com/document/big-data-shining-the-light-on-enterprise-dark-data.html

[26] 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

[27] Kevin, N. M., et. al. (2016). Dark data: Business Analytical tools and Facilities for illuminating dark data. Scientific Research Journal. 4, 1-10.

[28] Lehmann, H. (2010). The Dynamics of International Information Systems: Anatomy of a Grounded Theory Investigation. New Zealand: Springer.

[29] Liu, Y., et al. (2019). A Framework for Image Dark Data Assessment. 10.1007/978-3-030-26072-9_1.

[30] Lugmayr, A., Stockleben, B., Scheib, C., Mailaparampil, M.A. (2017). Cognitive big data: survey and review on big data research and its implications. What is really “new” in big data? Journal of Knowledge Management. 21(1), 197-212.

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

[32] Martin, E. J. (2016). Dark Data: Analyzing Unused and Ignored Information. econtentMag.com

[33] Neff, E.P. (2018). Dark data see the light. Lab Animal. 47(2), 45-48.

[34] Northwoods (2017). Dark data defined: what it means and why it's critical for child welfare. Dublin: Northwoods.

[35] Patton, M. Q., (2014), Qualitative Research & Evaluation Methods Integrating Theory and Practice. 4th Edition. Sage Publications, Thousand Oaks, CA

[36] Rao, V. (2018). Extracting dark data.. Accessed: Oct. 16, 2020. [On- line]. Available: https:// developer.ibm.com/ articles/ ba- data- becomes- knowledge- 3/

[37] 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

[38] SMECorp (2020). SME Definition. Retrieved from https://www.smecorp.gov.my/index.php/en/policies/2020-02-11-08-01-24/sme-definition

[39] Splunk (2019). The state of dark data. Retrieved on July 3rd, 2020 from https://www.splunk.com/en_us/form/the-state-of-dark-data.html

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

[41] Trajanov, D., Zdraveski, V., Stojanov, R. & Kocarev, L. (2018). Dark Data in Internet of Things (IoT): Challenges and Opportunities. Proceedings of the 7th Small Systems Simulation Symposium 2018, Niš, Serbia, 12th-14th February 2018.

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

[43] Veritas (2016). State of Information Governance: 2016 Report. California: Veritas

[44] Veritas, DLT Solutions & GovLoop (2017). Dark Data Management: The Next Frontier for Government Data. Washington: GovLoop