

**Computing Research Project**

**40201391**

**K/618/7425**

**Section (7)**

**Shedding Light on Dark Data: The Critical Role of the Scientific Data Officer**

**Submitted to**

Eng. Islam Alomari

Eng. Batool Alarmouti

**Submitted on**

March 10th, 2024

**Submitted by**

Marwan Tarek Shafiq Al Farah

**Student ID**

21110011

**Fall Semester 2023 - 2024**

**Table of Content**

[***Introduction*** 3](#_Toc159571673)

[***Literature Review*** 5](#_Toc159571674)

[***Research Questions and Objectives*** 7](#_Toc159571675)

[***Research Methodology*** 8](#_Toc159571676)

[***Conclusion*** 20](#_Toc159571677)

[***References*** 22](#_Toc159571678)

# ***Introduction***

Big Data has become an integral part of our information-driven age as data is the foundation to the development of the digital ecosystem. This concept transcends the realms of tech enthusiasts, capturing the attention of CEOs, healthcare professionals, and politicians. It is recognized as a transformation agent and an inventive spark. (Nguyen, 2019; Saeed and Husamaldin, 2021).

The 5 Vs, which stand for volume, velocity, variety, veracity, and value, are what define big data as they characterize its enormous scale, rapid generation and processing speed, variety of sources and formats, reliability, and crucial insights it produces (Nguyen, 2019; Saeed and Husamaldin, 2021). With the expectation that the digital universe will reach 175 zettabytes by 2025, the potential and problems associated with Big Data management are becoming more apparent (Tang, Ma and Luo, 2020). Its vast variety requires sophisticated algorithms and analytical tools for processing, as it comprises both structured and unstructured data (Saeed and Husamaldin, 2021; Vranopoulos, Clarke and Atkinson, 2022).

Big Data’s value comes from its ability to improve consumer experiences, stimulate innovation in a variety of industries, and inform decision-making. It helps with individualized care in healthcare and improves fraud detection and risk management in finance. Big Data has the ability to completely change a wide range of sectors and lead the way in new scientific and technological developments. (Faridoon and Imran, 2021)

Moving on from the larger effects of Big Data, one important point that becomes clear is the storage formats that enable this data revolution. Novel Big Data storage technologies have evolved when traditional storage options, such as relational databases, become insufficient due to their restricted scalability and flexibility. (Khan *et al.*, 2019; Faridoon and Imran, 2021)

Storage across several devices is made possible by distributed storage systems like HDFS (Khan *et al.*, 2019; Faridoon and Imran, 2021). Object storage systems, such as OpenStack Swift and Amazon S3, offer scalable designs that are perfect for managing enormous volumes of unstructured data (Bhagat, Tamang and Saxena, 2020). NoSQL databases like MongoDB and Cassandra handle the variety of Big Data without the confines of traditional SQL databases (Khan *et al.*, 2019; Faridoon and Imran, 2021).

These achievements, however, are not without challenges. Data access speed, security, and privacy compliance are all critical considerations. Furthermore, the desire to hoard data, fueled by a belief in its future value, leads to increased storage costs and concerns about efficiency. This emphasizes the significance of establishing storage models that prioritize not just efficiency and security, but also detecting and managing ROT (Redundant, Obsolete, or Trivial data) and Dark Data, which further complicate data management (Chant, 2023; George *et al.*, 2023). This is especially crucial in light of the fact that about 90% of all organizational data is unused, costing businesses a great deal of money and resources and making data management far more challenging than it has to be (George *et al.*, 2023). On the other hand, this data, while often overlooked, can hold great value. When properly used, it has the potential to yield a considerable Return on Investment from the dark data that remains untapped in a company’s database (George *et al.*, 2023).

This study aims to investigate the issue of data classification and its impact on organizational efficiency and data utilization. By delving into the categorization of data as either business critical, dark, or Redundant, Obsolete, and Trivial (ROT), this study seeks to evaluate effective methodologies for accurate and efficient data classification. Additionally, the research explores the potential of the Scientific Data Officer (SDO) role in optimizing data management, addressing the challenges associated with dark and ROT data, and unlocking their hidden value. The research aims to provide organizations with actionable insights and practical strategies for improving big data management practices, maximizing data utilization, lowering costs, and achieving greater operational efficiency.

# ***Literature Review***

In today’s data-driven world, organizations collect, process, and store vast amounts of data, however, a significant part of this data, known as dark data, remains untapped due to its unstructured, unorganized, or unknown nature. The concept of dark data is explored comprehensively in seven key papers. Each paper offers unique insights into the identification, management, and strategic utilization of dark data.

George et al. made a study that serves as a cornerstone in understanding dark data. The authors construct a robust framework for businesses to harness dark data. They introduce several challenges associated with dark data identification and value extraction then delve into methodologies for identifying and extracting valuable insights from this data, highlighting its capacity to transform decision-making processes. The authors also explore the potential of dark data in enhancing customer experiences and uncovering untapped market opportunities, reducing costs, and mitigating risks. They further discuss the technological and organizational challenges in integrating dark data, offering solutions for effective data governance and analytics (George *et al.*, 2023).

Schembera and Durán focused on dark data in high-performance computing environments in their paper, indicating that dark data forms 80% of the stored data in HPC facilities. They introduce the concept of the Scientific Data Officer (SDO) to manage dark data. They argue for the necessity of this role in ensuring systematic data management, which includes the discovery, organization, documentation, and preservation of dark data. The need for an SDO stems from the fact that a lack of documentation and metadata, as well as the challenge of data scattered across multiple storage systems, and as older data may become inaccessible due to outdated software and hardware, all lead to the creation of dark data, which presents a number of challenges such as inefficient resource utilization, hindered collaboration, and loss of potential knowledge. Their research illuminates how the effective use of dark data can foster better collaboration and knowledge discovery in scientific research, contributing to more meaningful scientific advancements (Schembera and Durán, 2020).

In their exploration of the Internet of Things (IoT), Chakrabarty and Joshi provided a comprehensive framework for managing dark data in IoT environments. They emphasize the need for real-time data processing and innovative recovery techniques while ensuring the identification and utilization of dark data. The authors proposed the use of soft sensors and the DeepDive framework, detailing how these tools can significantly improve the efficiency and functionality of IoT systems. Their research offers a promising perspective on the potential of well-managed dark data in revolutionizing IoT applications leading to enhanced decision-making and enhanced predictive analysis (Chakrabarty and Joshi, 2020).

Liu et al. addressed the challenge of managing dark data in digital imagery in their paper. The authors developed an advanced framework employing deep self-taught hashing (DSTH) and semantic hash ranking (SHR) for effective semantic feature extraction from images. This methodology is particularly groundbreaking in handling large-scale unstructured image datasets, providing a new avenue for data management and analysis in the field of digital imagery (Liu *et al.*, 2019).

The UK 2020 Databerg Report Revisited by Veritas sheds light on the complexities of managing dark and ROT (Redundant, Obsolete, Trivial) data in cloud environments. The report reflects on the challenges organizations face in adopting cloud technology, particularly concerning data management strategies. It emphasizes the importance of understanding and addressing the complexities associated with cloud-based dark and ROT data, as dark data accounts for approximately 53% of organizational data in the UK and ROT data accounts for approximately 29%, despite 97% of organizations having strategies for deleting ROT data, providing insights into best practices for data governance and compliance (Veritas, 2020).

Gimpel’s research is a deep dive into the invisible resource of dark IoT data within organizations. He comprehensively discussed the challenges, strategies to overcome these challenges, and strategies for leveraging this data, emphasizing the need to overcome organizational inertia. Gimpel proposes innovative approaches for integrating dark data into strategic decision-making, illustrating how this integration can lead to significant improvements in business performance and operational efficiency (Gimpel, 2020b).

In a separate but related work, Gimpel examined the strategic advantages of harnessing dark data in IoT. The author offered a detailed analysis of how underutilized IoT data can be a game-changer for businesses. The paper discussed the potential for transforming business processes, enhancing customer engagement, improving returns on assets, and driving product innovation through the strategic use of IoT dark data. Gimpel’s insights provide a roadmap for organizations looking to capitalize on the untapped potential of IoT data (Gimpel, 2020a).

In summary, these papers collectively highlight the vast and varied potential of dark data across different domains. While the challenges are substantial, including data identification, integration, and analysis, the opportunities for enhancing operational efficiency, customer experiences, and innovation are significant. Future research should focus on developing advanced dark data identification and management techniques, transforming dark data into usable and accessible formats, investigating how organizations can effectively manage and integrate dark data into their existing data systems, developing advanced methods for real-time dark data processing and analysis, designing educational programs to equip professionals with the skills required for effective dark data management, and explore the long-term impacts of dark data utilization on business and scientific research.

# ***Research Questions and Objectives***

**Research Questions:**

* What data classification frameworks can organizations effectively implement to accurately categorize data as business critical, dark, or ROT?
* How can the introduction of a Scientific Data Officer role contribute to improving data management, mitigating dark data challenges, and unlocking the potential value of underutilized data?

**Research Objectives:**

* Evaluate a framework or tool to assist in the data classification process and evaluate existing data classification methodologies and explore potential improvements for effective classification into business critical, dark, and ROT categories.
* Analyze the potential impact of the SDO role on data management practices, focusing on dark data identification, utilization, and risk mitigation.
* Investigate the challenges and opportunities associated with SDO implementation in different organizational structures and explore the potential benefits and drawbacks of employing the SDO role, considering factors such as cost, skill requirements, and impact on existing organizational structures.

# ***Research Methodology***

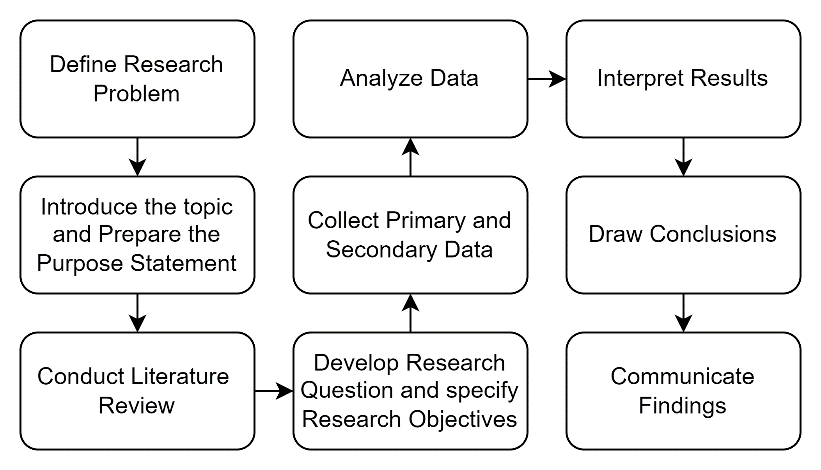
***Methodological Framework***

To understand the methodological framework of this study, we start with an introduction to Saunders’s research onion model, a comprehensive structure for delineating the layers of research design and methodology. This model guides the selection of appropriate research philosophy, approach, strategies, choices, time horizon, and data collection and analysis techniques. It offers a methodical way to handling the difficulties of looking into dark data management and the function of the Scientific Data Officer (SDO), making sure that every methodological decision is related to the others and forms a logical research plan.

1. **Research Philosophy: Interpretivism:** An interpretivist research philosophy is best appropriate for this study since it explores the roles, perspectives, and difficulties experienced by Scientific Data Officers (SDOs) in handling dark data. This philosophy supports the understanding of complex, context-dependent human behaviors and organizational roles. Interpretivism allows for a deep dive into the subjective experiences and interpretations of SDOs, acknowledging that reality is socially constructed and varies across individuals and settings.
2. **Research Approach: Qualitative:** The study takes a qualitative research technique, rooted in interpretivism, which is supported by the requirement to record in-depth, contextual insights into the strategies and experiences of SDOs. This method is perfect for investigating new or under-researched phenomena, such the growing significance of SDOs and the handling of dark data in businesses. By enabling the investigation of processes, meanings, and comprehensions that are difficult to quantify, qualitative approaches offer a deep, descriptive grasp of the study topic.
3. **Research Strategies: Case Studies and Thematic Analysis:** Employing case studies enables an in-depth examination of the SDO role within specific organizational contexts. This strategy is particularly effective for exploring contemporary phenomena within real-life contexts, where the boundaries between phenomenon and context are not clearly evident. Case studies provide the flexibility to use multiple sources of evidence, which is crucial for a comprehensive understanding of the SDO role and the management of dark data. Thematic analysis, as part of the qualitative analysis, is justified by its utility in identifying, analyzing, and reporting patterns (themes) within data. It allows for the organization and description of the data set in rich detail and interprets various aspects of the research topic. Both of these strategies build upon our research strategy and philosophy.
4. **Choice of Methods: Interviews and Document Analysis:** These methods are specifically chosen to synergize with the preceding layers. Semi-structured interviews are chosen for their flexibility, allowing for the exploration of the SDOs’ experiences, perceptions, and strategies in managing dark data, directly supporting the qualitative approach. This method facilitates a deeper understanding of the participants’ perspectives, providing insights that structured methods might miss. Document analysis complements interviews by providing additional context and background information from secondary data sources. Analyzing organizational documents, policies, and procedures related to data management offers a broader understanding of the environment in which SDOs operate, enhancing the depth of case studies and thematic analysis by adding context and background.
5. **Time Horizon: Cross-sectional:** A cross-sectional study is appropriate for this research, given the aim to explore the current state and perspectives of SDOs regarding dark data management. This design allows for the collection of data at a single point in time, offering a snapshot of practices and challenges across different organizations.
6. **Data Collection and Analysis:** Primary data collection through interviews and document analysis is directly informed by the study’s qualitative nature and ensures the relevance and specificity of the information to the research questions. The thematic analysis bridges all elements, from the interpretivist philosophy to the qualitative approach and case study strategy, will enable the identification of common themes and patterns regarding the role of SDOs in managing dark data, providing insights into effective practices and challenges.

The chosen layers and options, anchored in Saunders’s research onion model, are ideally suited for our investigation into dark data management and the pivotal role of the Scientific Data Officer (SDO). These methodologies offer a holistic and flexible framework, enabling a profound exploration of the nuanced and multifaceted nature of dark data within organizations. The interpretivist philosophy, qualitative approach, and employment of case studies and thematic analysis are particularly adept at uncovering the subjective experiences, organizational cultures, and contextual challenges faced by SDOs. This approach ensures that our research captures the depth, diversity, and complexity of SDO roles in navigating and leveraging dark data, thereby providing valuable insights into effective data management strategies and fostering a more informed understanding of the evolving landscape of organizational data practices.

***Research Flowchart***

****

**Figure 1 – Research Flowchart**

A structured research methodology was employed to explore the intricate role of data management. This section evaluates the research methodologies and processes applied to the project and justifies the chosen methods and analysis.

* **Defining the Research Problem:** The initial phase involved a precise definition of the research problem. The complex nature of dark data requires an approach that balances technical, organizational, and strategic considerations. A clear articulation of this problem set the foundation for the research, guiding subsequent methodological choices.
* **Purpose Statement and Literature Review:** A comprehensive literature review was conducted, which ensured a robust conceptual framework. This review mapped the existing knowledge landscape, identified gaps, and solidified the project’s purpose statement. This phase was critical in establishing a baseline of the dark data phenomenon and the nascent role of the Scientific Data Officer (SDO).
* **Research Question and Objectives:** Based on the literature insights, a research question was formulated along with specific objectives. These were designed to investigate how SDOs can illuminate and leverage dark data, thus driving value for organizations. The question was intentionally crafted to be open-ended, promoting exploratory research that is well-suited for emerging technological fields.
* **Data Collection and Analysis:** A mixed-methods approach was chosen for data collection, combining qualitative insights from expert interviews with qualitative and quantitative data from the literature review. This made it possible to see the issue from several angles and to see both subtle and statistical patterns. Iterative deductive and inductive reasoning were used in the data analysis process, allowing for a thorough examination of the themes and patterns that emerged.
* **Interpretation and Conclusion:** The interpretive phase was iterative, with findings continuously referenced against the research objectives to ensure alignment. The conclusions drawn were both inferential, providing generalizations about the role of SDOs, and indicative, suggesting potential strategies for managing dark data.
* **Communication of Findings:** Finally, the communication of findings can be conducted through academic channels and industry reports, ensuring both rigor and relevance. The flow of information from definition to dissemination was seamless, reflective of a well-orchestrated research methodology.
* **Primary Research Methodology:** As for the research methodologies, the qualitative research methodology was selected for our primary research and is well-justified given the exploratory nature of the study. Qualitative research is instrumental in understanding complex concepts and behaviors, particularly in new fields of study where not much has been established in terms of quantitative data. Qualitative methods are suitable for this research due to the need to capture detailed insights into the roles, perceptions, and challenges faced by Scientific Data Officers (SDOs) in managing dark data. Through interviews, case studies, and thematic analysis, the research captures the nuanced practices and experiences that quantitative data could overlook.
* **Secondary Research Methods:** The chosen research methods were justified on the grounds of their appropriateness for the exploratory nature of the project, the complexity of the subject matter, and the need for comprehensive, multi-dimensional insights. The mixed-methods approach provided a balance between breadth and depth, while the structured flow from literature review to communication ensured methodological rigor and clarity of findings.
* **Analysis Justification:** The thematic analysis enables the identification of patterns and themes within the qualitative data, providing a deep understanding of the underlying issues in dark data management. This approach aligns with the project’s goal to illuminate the often-invisible struggles and strategies of SDOs, thereby contributing to the body of knowledge with rich, contextualized findings.

The research methodologies and processes used were well thought out to address the many facets of the SDO’s responsibilities. The methodical technique made sure that every stage of the study was influenced by the results of the earlier one, resulting in a well-organized and perceptive examination of the topic. The study fully tackles the complexity of dark data in the field of computer research thanks to its qualitative methodology and careful research procedure.

***Data Collection and Analysis***

Primary data is data that the researcher collects directly for the purpose of their study. It involves collecting firsthand information using a variety of approaches, including surveys, interviews, experiments, and observations. This form of data is original and obtained directly from the source, offering specific insights into the study subject.

In order to gather primary data for our study, data management specialists were interviewed. This allowed us to get firsthand knowledge of the difficulties and solutions related to handling dark and ROT data. This methodology allowed for a thorough examination of the practical aspects of data management positions, namely the job of the Scientific Data Officer across various organizational settings.

Secondary data is data that was collected by someone else for a different reason but is then used by a researcher for their own study. This may contain data from past study, government records, historical documents, and data gathered by organizations. Secondary data is pre-existing and can provide additional context or support main data conclusions.

Secondary data were sourced from an extensive literature review, encompassing existing research, academic papers, and theoretical frameworks related to data management, classification, and the utilization of dark data. This review served as the foundation for understanding the broader academic and practical implications of our research findings.

**Access and Ethical Issues**

During the data collection phase, navigating access and ethical issues was critical to maintaining the integrity of our research. Access challenges included securing participation from a diverse range of data management professionals, which required extensive outreach and coordination. Ensuring the confidentiality of information shared by interviewees was paramount, involving stringent data handling protocols to anonymize responses and safeguard participant identities.

Informed consent was a cornerstone of our ethical approach, entailing clear communication about the study’s purpose, how the data would be used, and participants’ rights, including withdrawal. In order to build openness and confidence with our participants, this approach was essential. Ethical issues included protecting participant privacy, and taking measures to guarantee that all gathered information could not be linked back to specific participants.

Despite the robust ethical framework, the data collection process was not without its limitations. The possibility of selection bias was a worry, as the participants who were interviewed may not reflect the whole population of data management experts. Furthermore, the difficulty of generalizing findings from a small, potentially unrepresentative sample of interviews to larger groups or different settings was taken into consideration. These limitations highlight the importance of carefully interpreting the findings, taking into account the unique situations from which the data were gathered.

**Merits**

However, the merits of our approach were significant. The ability to obtain detailed, context-specific insights directly from data management experts provided a rich and nuanced understanding of the challenges and strategies in managing dark and ROT data. This firsthand knowledge significantly enhanced the applicability and depth of our research findings, contributing valuable perspectives to the discourse on data management practices.

The research on the Scientific Data Officer (SDO) role significantly enhances our understanding and execution of data management practices. It highlights the essential need for dedicated positions within organizations to oversee, classify, and utilize dark data. Through comprehensive literature reviews, case studies, and expert interviews, the study showcases how SDOs can bridge the gap between data collection and data utilization, leading to more informed decision-making and maximizing the value of untapped data resources.

Furthermore, the research offers actionable recommendations for organizations, advocating for the institutionalization of the SDO role and the adoption of advanced data management frameworks. The study provides businesses seeking to improve their data-driven capabilities with essential guidance by addressing the difficulties associated with managing dark data and highlighting the possible advantages of good data stewardship activities. It emphasizes how crucial SDOs are to creating an innovative and continuous improvement culture through strategic data management, which eventually boosts organizational competitiveness and efficiency.

**Limitations**

This research provides valuable insights into Big Data management and the role of the SDO, yet it’s crucial to recognize its limitations for a comprehensive understanding:

1. **Sample Size and Geographic Focus:** The study’s insights are drawn from interviews with professionals in just two organizations within Jordan, limiting the diversity of perspectives and practices. This narrow focus may not fully capture the varied challenges and strategies across different industries and geographic locations. Future research should aim for a more extensive and varied sample to enhance the generalizability of the findings to other contexts and organizational structures.
2. **Temporal and Methodological Constraints:** The research was conducted over a short period, preventing a longitudinal analysis of the SDO role’s impact. Additionally, the reliance on semi-structured interviews introduces potential biases, including selection bias and subjective interpretations. These factors could skew the findings, emphasizing the need for mixed-method approaches and longer-term studies to validate and deepen the understanding of the SDO role’s effectiveness and challenges.
3. **Scope of Recommendations and Rapid Technological Evolution:** The recommendations focus on adopting the SDO role and implementing advanced data classification frameworks at the organizational level, without fully exploring the broader implications for data governance, security, and ethical management of sensitive data. Moreover, the rapid evolution of data management technologies might quickly outdate the study’s findings, underscoring the importance of continuous research to adapt strategies to the latest advancements.
4. **Generalization of Findings:** The study’s specific focus on managing dark and ROT data within the context of the SDO role means its findings may not apply universally across all aspects of data management or to organizations with different digital footprints. The insights and recommendations should be applied with caution, considering the unique challenges and opportunities of each organizational and technological context.

In summary, while the study sheds light on managing dark and ROT data and the potential benefits of the SDO role, these limitations highlight the necessity for broader, more diverse, and methodologically varied research to fully understand and leverage the SDO role in data management across different industries and settings.

***Data Analysis***

Semi-structured interviews were the main method used to gather primary data as they allowed for in-depth conversations and the freedom to investigate the new role of the SDO and other data management practices. Interviews are a qualitative research approach that use structured or semi-structured questions, that offers deep insights into the practices, experiences, and difficulties associated with handling dark and ROT data. Interviews, as our primary data collection tool, facilitated a nuanced dialogue with professionals, employing a semi-structured format that balanced consistency with the flexibility to explore emergent themes. This method proved invaluable in capturing the complex realities of data management practices.

Thematic analysis was used to find, evaluate, and report on patterns (themes) in the data by categorizing qualitative data, organizing codes into relevant themes, and evaluating the relevance of those themes to the research objectives. Thematic analysis enabled us to reduce the complexity of the information gathered from interviews and turn them into actionable insights, emphasizing the critical role of Scientific Data Officers in tackling the diverse difficulties connected with dark data and offering ways for successful management.

**Primary Data Analysis**

***First Interview:***

In a discussion with a Senior Data Science & AI Specialist - Officer at bank in Jordan, he shared his insights on the proposed role of the Scientific Data Officer (SDO) and the evolving landscape of data management. His perspectives shed light on the complexities of handling unstructured data and the strategic importance of the SDO role in modern organizations.

He perceives the SDO as a new yet vital position within organizations, particularly due to the increasing challenge of managing ROT (Redundant, Obsolete, Trivial) and Dark data. As businesses delve deeper into advanced techniques like Large Language Models (LLMs) and multi-model systems, they encounter a pressing need to harness unstructured data, which was previously overlooked or deemed too complex to process. He emphasized that the SDO could play a crucial role in bridging this gap by working across departments to identify and capture necessary data, thus serving as a pivotal point of contact for data-related inquiries and strategies, as the SDO has the responsibility to meet with every department, identify their projects, and identify what data needs to be captured and what data they already have.

He discussed the bank’s recent shift towards exploring previously uncaptured data sources, driven by the realization of their potential value and the advancements in technology that now allow for effective management of unstructured data. He acknowledged that such data were often neglected due to perceived irrelevance, processing limitations, or the sheer complexity of structuring it. He pointed out the need for reevaluating how data is captured and utilized, citing instances where critical information is either inadequately summarized or lost, which can significantly impact decision-making and operational efficiency.

Looking ahead, he anticipates that the SDO role will not only become more prevalent but also essential in most large organizations in the near future. As the volume of unstructured data grows exponentially, the challenges it poses will necessitate a dedicated role focusing on its management. He suggests that while some organizations might designate a specific individual as an SDO, others might integrate these responsibilities into the roles of existing data scientists.

He also expressed a strong conviction in the practicality and relevance of the research on the Scientific Data Officer (SDO) role, particularly for their Bank, which is currently recognizing the need for such a position to comprehensively manage its data. He emphasized the significance of examining how the introduction of the SDO might transform data capture and utilization pipelines. He noted that some of the current processes often fail to capture the full context or details of data, as it not being recorded properly or that it is not being comprehensive enough. He cited an example where summaries of conversations might inadequately reflect the issue resolved, the solutions provided, and the time taken. He gave an example of simply recording “problem solved” rather that explaining what the actual problem was or what it solution was. He suggested that ensuring well-captured dataset that could reveal insights previously unanticipated by the organization should be one of the responsibilities of the SDO.

He proposed expanding the SDO’s responsibilities to include identifying new types of data to capture, determining the methods for doing so, and suggesting changes to the current data processing pipelines and storage formats. He acknowledged that while he had not previously heard of the SDO role, he anticipated the emergence of such a position due to the existing gaps in data management and the advent of new roles brought about by technological advancements and requirements, such as prompt engineering and AI engineering.

He envisages the primary objective of the SDO to be managing dark data, which he believes will significantly enhance data-driven decision-making. This role, according to him, is crucial in establishing the foundations for new technologies that will address more complex problems as it might help identify some problems in the organization that will lead the organization to make more data driven decisions.

However, he noted the challenges in managing metadata, which often arise from formatting issues or time constraints as there is no way to capture everything. He stressed that it’s the SDO’s responsibility to reformulate how metadata is captured to ensure it’s both comprehensive and timely.

To convince departments and management of the value of capturing specific data types, he highlighted the importance of demonstrating their utility, perhaps through a proof of concept or a simple dashboard, to garner support and understanding for these initiatives. He shared an anecdote where his department, initially overlooked, became inundated with tasks once they demonstrated the added value they could provide. This narrative underscores the importance of showing practical benefits to secure organizational buy-in for new data strategies.

The introduction of APARAVI’s Intelligent Management of Unstructured Data captivated his interest, particularly from the perspective of enhancing the SDO’s toolkit. Unfamiliar with the platform before the interview, he quickly recognized its potential to revolutionize the SDO’s approach to managing the bank’s burgeoning unstructured data. He saw APARAVI as a powerful ally for the SDO, offering the means to effectively navigate and curate vast amounts of unstructured data, thereby reducing the prevalence of dark data. The platform’s promise to streamline data lifecycle processes in line with FAIR principles aligned perfectly with his vision of the SDO’s role in reaching robust, ethical, and efficient data management strategies. He envisaged the SDO utilizing APARAVI to not only capture but also maximize the utility of valuable information, transforming the way the bank leverages its data assets.

His interest was equally piqued by the concept of New Classification System Development, seeing it as a strategic enhancement to the SDO’s capabilities in managing data. He recognized the system’s potential to transform the SDO’s approach to data management by enabling proactive categorization of new data during storage by all employees, distinguishing between business-critical and potentially dark data. This foresight, combined with the ability to assign retention periods to ROT data, appeared to him as an effective strategy for the SDO to tackle the challenges of dark data. He appreciated how this system would empower the SDO to make informed decisions on data archiving or deletion, optimizing storage, bolstering security, and ensuring compliance.

In conclusion, his reflections reveal a clear understanding of the need for the SDO role, specifically in companies that have a CDO, such as their bank, and potentially across the industry. His anticipation of the role’s evolution and his emphasis on the strategic need to expand and enhance data management practices reflect a forward-thinking approach to the burgeoning field of data science and its implications for organizational success and innovation.

***Second Interview:***

In an interview with the Intelligence and Analytics Unit Manager at a software company in Jordan, we delved into the complexities and potential of managing dark data and the emerging role of the Scientific Data Officer (SDO). He shares his experiences and perspectives on the significance of these roles in enhancing data-driven strategies and organizational efficiency. His expert insights offer a deeper understanding of the challenges and opportunities within the realm of data management.

He demonstrated a general knowledge and previous understanding of ROT (Redundant, Obsolete, or Trivial data) and Dark Data. He highlighted that dark data is prevalent not only in unstructured data but also within structured data. His insights revealed that when previously unused data was leveraged, it significantly benefited their organization by enhancing marketing strategies, increasing profits, reaching new market segments, and acquiring new customers.

He regards the Scientific Data Officer (SDO) as a crucial and emerging role within the data management sector that will pose many challenges but will reflect various profits to the organization.

From his experience, he confirmed that dark data presents a significant issue, with his organization frequently encountering situations where data is captured and stored without being utilized. He noted that although the organization is not primarily a big data company, it extensively interacts with big data organizations, including telecom operators and major insurance firms, highlighting the pervasive nature of dark data across industries.

He mentioned that his organization recently initiated an AI team but has yet to appoint a Chief Data Officer. He anticipates that the SDO role will become a standard in IT companies, paralleling the evolutionary trajectory of AI roles, which gained prominence and necessity after demonstrating tangible value in the industry. He believes that the SDO role will similarly become indispensable once its value is widely recognized and validated.

He emphasized that an effective SDO must possess a profound understanding of the company’s data, alongside excellent judgment skills to determine the data’s importance, potential value, and utility. He noted that the role extends beyond individual efforts, requiring collaborative synergy across various departments to fully harness its potential. He stressed the importance of strong communication skills for an SDO, as articulating the rationale and value behind data-related decisions is key to successful implementation and organizational buy-in.

He also believes that the SDO should have a specific set of skills that will help him in doing these roles. He outlined a comprehensive skill set for an effective SDO, including data management, critical thinking, exploratory skills, and a holistic understanding of the company’s data landscape. He emphasized the need for initiative, presentation and persuasion abilities, a willingness to improve the self, and the capacity to structure unstructured data. Additionally, he noted that general professional skills such as time management, organization, and strong social skills are equally important for the SDO role.

He was intrigued by APARAVI, a platform previously unknown to him, which he envisioned could empower the role of the Scientific Data Officer (SDO) within their organization. Recognizing its potential to radically transform the SDO’s approach to managing, categorizing, and leveraging unstructured and dark data, he saw APARAVI as an invaluable tool in the SDO’s arsenal. The platform’s potential to help with dark data identification and classification as well as offer practical methods for utilizing or disposing of it made a strong argument for its incorporation into their data management plan. He contemplated how the SDO, equipped with APARAVI, could drive the organization towards heightened data transparency, operational efficiency, and strategic depth. Despite his enthusiasm, he maintained a cautious stance, advocating for a comprehensive evaluation to ensure the platform’s alignment with the unique data environment and goals of the organization.

Additionally, he expressed interest in the potential impact of a New Classification System Development on the role of the SDO in their organization. He recognized the proactive classification of new data as either business-critical or potentially dark during storage as an innovative approach to enhance the SDO’s data management strategy. This system’s ability to assign retention periods to dark or ROT data, followed by deletion or archiving, resonated with him as a logical extension of the SDO’s responsibilities in maintaining data hygiene and operational efficiency. He foresaw how this system could not only alleviate the burden of managing increasing data volumes but also align with the SDO’s mission to optimize storage, mitigate risks, and unlock the hidden value in their data assets. Acknowledging the inherent challenges of implementing such a system, including the need for meticulous data assessment and potential cultural resistance, he was nonetheless eager to see how it could be customized to enhance the SDO’s capability to orchestrate a more strategic, streamlined, and effective data management framework.

His insights provide a valuable perspective on the significance of managing dark data and the emerging role of the SDO. His emphasis on collaboration, communication, and a diverse skill set underscores the multifaceted nature of the SDO role and its potential impact on data-driven decision-making and organizational efficiency. As their organization navigates its journey in AI and data management, the insights from this interview reflect broader industry trends and the evolving landscape of data roles within organizations.

**Secondary Data Analysis**

The secondary data analysis conducted through a comprehensive literature review reveals pivotal insights into the realm of dark data, its management, and strategic utilization within various sectors. This review accentuates the untapped potential of dark data, which remains largely unexplored due to its unstructured, unorganized, or unknown nature. The synthesis of findings across seven key papers elucidates a spectrum of methodologies, challenges, and the transformative impact that effective dark data management can have on decision-making processes, customer experiences, operational efficiency, and market opportunities.

The cornerstone study by George et al. (2023) lays the groundwork for understanding dark data, offering a robust framework for businesses to harness its latent value. This study, along with others in the field such as Schembera and Durán (2020) and Gimpel (2020a, 2020b), underscores the necessity of the Scientific Data Officer (SDO) role in navigating the complexities of dark data since they reveal that dark data could form up to 80% of stored data, pointing towards a vast unexplored asset. The UK 2020 Databerg Report revisited by Veritas (2020) indicates that dark data constitutes about 53% of organizational data, with ROT data making up another 29%. These papers collectively highlight the significance of adopting innovative technologies, frameworks like DeepDive, and strategic approaches to unlock the hidden potential of dark data. Notably, the integration of such data can foster enhanced decision-making, optimize resource utilization, and drive scientific and technological advancements. The challenge of dark data sprawled across high-performance computing environments and IoT platforms calls for specialized roles and sophisticated management strategies to ensure its discovery, organization, documentation, and preservation.

In summary, the literature review not only spotlights the critical issues associated with managing dark data but also proposes actionable insights for organizations to reevaluate their data management strategies. By embracing the concept of the Scientific Data Officer and leveraging advanced data classification frameworks, businesses can significantly mitigate the challenges posed by dark and ROT data. The journey towards effective dark data management is complex yet rewarding, offering a path to operational excellence and a competitive edge in today’s data-driven landscape. Future research directions should focus on refining dark data identification and management techniques, fostering a deeper integration of dark data into organizational data systems, and exploring the long-term impacts of such initiatives on business innovation and scientific research.

**Results and Findings**

The research conducted has yielded significant findings. Through literature review and primary data collection via interviews with professionals in the field, the study explored the challenges and opportunities of managing Big Data, with a focus on dark data and ROT (Redundant, Obsolete, or Trivial) data. The introduction of the SDO role as a strategic solution to these challenges was a central theme.

***Key Findings from Literature Review***

The literature review highlighted the pervasive issue of dark data within organizations, identifying it as a largely untapped resource due to its unstructured, unorganized, or unknown nature. The reviewed papers collectively emphasized the potential of dark data in transforming decision-making processes, enhancing customer experiences, and uncovering new market opportunities. Furthermore, the concept of the Scientific Data Officer (SDO) emerged as a necessary role for systematic data management, emphasizing the need for a dedicated professional to handle the complexities of dark and ROT data.

***Insights from Interviews***

Interviews with data professionals from a bank and a software company in Jordan provided practical insights into the current state of data management and the perceived benefits of the SDO role:

* **Senior Data Science & AI Specialist - Officer at a bank in Jordan:** This interviewee recognized the SDO role as vital for bridging gaps in managing ROT and dark data. The discussion highlighted the bank’s shift towards leveraging previously uncaptured data, driven by technological advancements that facilitate the management of unstructured data. The interviewee anticipated the SDO role becoming essential in large organizations to tackle the growing volume of unstructured data.
* **Intelligence and Analytics Unit Manager at a software company in Jordan:** This interview revealed a general awareness of the challenges posed by ROT and dark data, even within structured data environments. The potential benefits of leveraging dark data for enhancing marketing strategies and customer acquisition were emphasized. The interviewee also acknowledged the necessity of the SDO role in orchestrating effective data management strategies, highlighting the importance of a comprehensive skill set including data management, critical thinking, and strong communication.

**Alignment with Research Objectives and Answers to Research Questions**

***Research Objectives:***

* **Evaluate a framework or tool to assist in the data classification process:** The findings from both literature and interviews underscore the complexity of managing dark and ROT data and the potential of the SDO role in addressing these challenges. The insights align with the objective of developing effective data classification methodologies, suggesting a need for tools that can aid in the accurate categorization of data. The interest in APARAVI and the concept of New Classification System Development from interviewees indicates a move towards implementing such frameworks.
* **Analyze the potential impact of the SDO role on data management practices:** The discussions with industry professionals validate the research objective of evaluating the SDO’s impact. The anticipation of the SDO role becoming a standard in IT companies and its strategic importance in managing dark data illustrate the role’s potential to significantly improve data management practices.
* **Investigate the challenges, opportunities, benefits, and drawbacks of employing the SDO role:** Introducing a Scientific Data Officer (SDO) role presents both challenges, such as integration into existing hierarchies, skill requirements, and cost implications, and opportunities, including enhanced data management, unlocking value from dark data, and driving organizational change. The SDO role’s impact on organizational structures requires careful consideration of governance, skills, and costs, yet offers significant potential for improved data utilization and operational efficiency. Organizations considering the SDO role must weigh these factors to effectively tailor the role to their specific needs.

***Answers to Research Questions:***

* **What data classification frameworks can organizations effectively implement to accurately categorize data as business critical, dark, or ROT?:** The research indicates that organizations are on the brink of adopting advanced data classification frameworks, with technologies like APARAVI being considered for their potential to assist in the categorization process. The development of new classification systems that enable categorization of data is seen as a strategic enhancement to the SDO’s capabilities.
* **How can the introduction of a Scientific Data Officer role contribute to improving data management, mitigating dark data challenges, and unlocking the potential value of underutilized data?:** The findings suggest that the SDO role is critical in managing the complexities of dark and ROT data. The role is expected to bridge departmental gaps, ensuring that data is captured comprehensively and utilized effectively. The insights from industry professionals underscore the SDO’s potential to drive data-driven decision-making and operational efficiency.

# ***Conclusion***

Based on the comprehensive analysis and findings of this study, several key recommendations are presented to enhance organizational efficiency and data utilization through effective data management practices. These recommendations are tailored to address the challenges associated with managing Big Data, particularly dark data, and ROT (Redundant, Obsolete, or Trivial) data, and to leverage the strategic value of the Scientific Data Officer (SDO) role.

1. **Institutionalize the Role of the Scientific Data Officer (SDO):** Companies have to think about making the SDO a key player in their data management strategy. Their role is essential for bridging the gaps between different data sources and guaranteeing thorough data collection, categorization, and usage. All departments should collaborate closely with the SDO to find, gather, and use data that can improve operational effectiveness and decision-making. This is because the insights gleaned from the interviews underscore the growing importance of the SDO in handling the complexity associated with big data. A committed professional may make sure that data management techniques are in line with corporate objectives and technical improvements as data volumes, velocities, and varieties continue to increase.
2. **Implement Advanced Data Classification Frameworks:** Organizations must adopt and implement advanced data classification frameworks that can accurately categorize data into business-critical, dark, or ROT categories. Tools like APARAVI and concepts such as the New Classification System Development should be explored for their potential to enhance the SDO’s capabilities in data management. This is due to the fact that the study’s findings suggest that effective data classification is foundational to optimizing data management practices. Advanced frameworks can aid in the accurate categorization of data, thereby improving data utilization, reducing storage costs, and ensuring compliance with data protection regulations.
3. **Leverage Technology for Enhanced Data Management:** Investment in technologies and tools that support the SDO in managing and analyzing Big Data is recommended. This includes tools for semantic feature extraction, deep learning algorithms for unstructured data analysis, and platforms for streamlined data lifecycle processes. This is due to the fact that the practical insights into technologies like APARAVI indicate the transformative potential of such tools in managing unstructured data. Leveraging technology can significantly reduce the prevalence of dark data, enhance operational efficiency, and unlock the hidden value in data assets.
4. **Foster a Data-Driven Organizational Culture:** Organizations should foster a data-driven culture that values and utilizes data as a strategic asset. This involves training and empowering employees across all levels to recognize the importance of data, understand basic data management practices, and contribute to the organization’s data classification efforts. This is due to the fact that the challenges of managing dark and ROT data are not only technological but also cultural. By fostering a data-driven culture, organizations can enhance the understanding and appreciation of data’s value, thereby supporting the SDO’s efforts in data management and utilization.
5. **Develop Continuous Learning and Improvement Mechanisms:** To ensure that the SDO and data management teams are up to date on the most recent advancements in data science, Big Data technologies, and data governance procedures, methods for continuous learning should be put in place. Regular training sessions, workshops, and involvement in industry forums are a few examples of this. This may be attributed to the fact that data management is a quickly changing sector where new technology and difficulties are always developing. The SDO and their team may be guaranteed to have the most up-to-date information and expertise to handle Big Data efficiently with the help of mechanisms for continuous learning and improvement.

# ***References***

Bhagat, V., Tamang, A. and Saxena, A. (2020) ‘Object-Based Storage Accompanying File System For Improved Data Persistent And Processing Capabilities’, *International Journal of Scientific & Technology Research*, 9(February 2021), p. 3601.

Chakrabarty, S. and Joshi, R.S. (2020) ‘Dark Data: People to People Recovery’, *Lecture Notes in Networks and Systems*, 93(January 2020), pp. 247–254. Available at: https://doi.org/10.1007/978-981-15-0630-7\_24.

Chant, G.G. (2023) ‘Dealing with Dark Data – Shining a Light’, *Communications in Computer and Information Science*, 1825 CCIS, pp. 149–160. Available at: https://doi.org/10.1007/978-3-031-34045-1\_13/COVER.

Faridoon, A. and Imran, M. (2021) ‘Big Data Storage Tools Using NoSQL Databases and Their Applications in Various Domains: A Systematic Review’, *Computing and Informatics* [Preprint].

George, A.S. *et al.* (2023) ‘Partners Universal International Innovation Journal (PUIIJ) Bringing Light to Dark Data: A Framework for Unlocking Hidden Business Value’, (August), pp. 35–60. Available at: https://doi.org/10.5281/zenodo.8262384.

Gimpel, G. (2020a) ‘Bringing dark data into the light: Illuminating existing IoT data lost within your organization’, *Business Horizons*, 63(4), pp. 519–530. Available at: https://doi.org/10.1016/j.bushor.2020.03.009.

Gimpel, G. (2020b) ‘Dark data: the invisible resource that can drive performance now’, *Journal of Business Strategy*, 42(4), pp. 223–232. Available at: https://doi.org/10.1108/JBS-02-2020-0046.

Khan, S. *et al.* (2019) ‘Storage Solutions for Big Data Systems: A Qualitative Study and Comparison’. Available at: http://arxiv.org/abs/1904.11498.

Liu, Yu *et al.* (2019) *A Framework for Image Dark Data*. Springer International Publishing. Available at: https://doi.org/10.1007/978-3-030-26072-9.

Nguyen, T.L. (2019) ‘A Framework for Five Big V’s of Big Data and Organizational Culture in Firms’, *Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018*, pp. 5411–5413. Available at: https://doi.org/10.1109/BigData.2018.8622377.

Saeed, N. and Husamaldin, L. (2021) ‘Big Data Characteristics (V’s) in Industry’, *Iraqi Journal of Industrial Research*, 8(1), pp. 1–9. Available at: https://doi.org/10.53523/ijoirvol8i1id52.

Schembera, B. and Durán, J.M. (2020) ‘Dark Data as the New Challenge for Big Data Science and the Introduction of the Scientific Data Officer’, *Philosophy and Technology*, 33(1), pp. 93–115. Available at: https://doi.org/10.1007/s13347-019-00346-x.

Tang, J., Ma, T. and Luo, Q. (2020) ‘Trends Prediction of Big Data：A Case Study based on Fusion Data’, *Procedia Computer Science*, pp. 181–190. Available at: https://doi.org/10.1016/j.procs.2020.06.073.

Veritas (2020) ‘The UK 2020 Databerg Report Revisited’, *Veritas* [Preprint].

Vranopoulos, G., Clarke, N. and Atkinson, S. (2022) ‘Addressing big data variety using an automated approach for data characterization’, *Journal of Big Data*, 9(1). Available at: https://doi.org/10.1186/s40537-021-00554-3.