University of Windsor

COMP 8157 Advanced Database Topics

Phase 03 - Final Report - submitted by Group 01 on Mar 26th, 2025

Group Name: Big Hero 06

Project Title: Does Big Data really need to be Big?

Contribution:

Student ID	110185281	110166390	110167100	110167885	110174849	110160683
Student Name	Arshnoor Singh Sohi	Ashiqur Rahman	Gagandeep Kaur .	Foysal Rahman Nitu	Tausif Zaman	Yingying Lin
Contribution	16.667	16.667	16.667	16.667	16.667	16.667
Signature	Arshnoor Singh Sohi	Ashiqur Rahman	Gagandeep Kaur	Foysal Rahman Nitu	Tausif Zaman	Yingying Lin

CONFIDENTIALITY AGREEMENT & STATEMENT OF HONESTY

We, members of Group 01, verify that the submitted work is our own, original work, that all sources are cited accurately, and that we have not submitted any portion of this work for any other university course. We did not use Generative AI tools (e.g., ChatGPT, Bard) to produce this assignment or report. We confirm knowing that a mark of 0 may be assigned for copied work.

Arshnoor Singh Sohi Student Signature	**Arshnoor Singh Sohi** Student Name (please print)	**110185281** Student I.D. Number
Ashiqur Rahman Student Signature	**Ashiqur Rahman** Student Name (please print)	**110166390** Student I.D. Number
Gagandeep Kaur Student Signature	**Gagandeep Kaur .** Student Name (please print)	**110167100** Student I.D. Number
Foysal Rahman Nitu Student Signature	**Foysal Rahman Nitu** Student Name (please print)	**110167885** Student I.D. Number
Tausif Zaman Student Signature	**Tausif Zaman** Student Name (please print)	**110174849** Student I.D. Number
Yingying Lin Student Signature	**Yingying Lin** Student Name (please print)	**110160683** Student I.D. Number

Does Big Data Really Need to Be Big? (Probably!?)

Arshnoor Singh Sohi
Department of Computer Science
University of Windsor
sohi21@uwindsor.ca

Ashiqur Rahman
Department of Computer Science
University of Windsor
rahman6s@uwindsor.ca

Gagandeep Kaur .
Department of Computer Science
University of Windsor
gagande3@uwindsor.ca

Foysal Rahman Nitu
Department of Computer Science
University of Windsor
nituf@uwindsor.ca

Tausif Zaman
Department of Computer Science
University of Windsor
zaman45@uwindsor.ca

Yingying Lin
Department of Computer Science
University of Windsor
lin8h@uwindsor.ca

***Abstract*— Does big data really need to be big? Turns out the scale is even larger, and moreover, much of it remains dark. Researchers suggest that at least 50% of collected data goes dark, with up to 90% of datasets in companies already classified as dark data. One major factor behind this accumulation is data hoarding by organizations fearing potential losses in the long run. This raises an important question: does dark data hold any long-term potential? Can untapped data be leveraged for present benefits? Evidence suggests it can.**

Dark data has gained significant attention from researchers in recent years, with efforts focused on harnessing the hidden value it possesses. The Dark Data Management Repercussion model explores the potential of seemingly redundant data. Existing frameworks, such as the FAIR standard and metadata approaches, attempt to address the issue but fail to provide a complete solution. To tackle this, the Quality-Driven Data Reduction (QDDR) framework is proposed, addressing data volume, utility, and analytics efficiency. The development of the Dark Data Transformation Module (DDTM) demonstrates a systematic method for converting disparate data formats into analyzable structures. Using synthetic healthcare data, the complexity of dark data management is highlighted, and methodologies are proposed for extracting value from previously overlooked information resources.

***Index Terms* — Big Data Management, Dark Data Transformation, Data Quality Assessment, Data Reduction Strategies, Machine Learning Data Optimization, Metadata Enhancement, Data Lifecycle Management, Anomaly Detection, Data Value Optimization, Enterprise Data Analytics, Synthetic Data Exploration, Data Format Conversion**

I. INTRODUCTION

In Computer Science topics such as blockchain, real-time analytics, anomaly detection, event prediction, all have one thing in common, the use of big data. Big data conventionally is thought of as something that requires "an enormous amount of data", in this paper aims to challenge that and analyze if its vastness is essential.

In this paper the following questions are posed, based on what one might stumble upon while thinking about the importance, quality, and usage of big data,

1) What amount of data is really needed for analytics?
2) Is it possible that **billions** of entries collected and stored are of no use?
3) Can the same analytical result be achieved by reducing the dataset?

Historically there have been different interpretations of big data used in the industry. The simplest of these considers the three V's, namely volume, velocity and variety [1] [2].

1) Volume – describes the huge amount of data being generated from numerous devices
2) Velocity – refers to the speed at which data is being generated
3) Variety – describes the heterogeneous nature of data being generated in different formats such as text, images and videos

Over time additional V's have been associated with big data. These include veracity, value, variability and visibility [2] [3]. For the most part, this paper will stick with the three original V's of big data, however, such as value of the dataset may be referred to for experimental purposes.

The advancement in hardware and storage systems can be accredited for the ability to collect massive amounts of data. Storage systems have evolved drastically over the last few decades to meet customer needs [4]. This has made it easier for humans to generate and store data without having to worry about storing it anymore. Moreover, based on a recent study [2], 147 zettabytes of data were generated in the year 2024 alone, 181 zettabytes of data will be generated in the year 2025 and it is only expected to grow. There has been a significant increase in the number of connected IoT devices as well [2], which contributes to a major share of big data. However, rapid growth of data has made tasks such as utilization and management of data difficult.

One of the reasons is the consumption of data, a significant amount of data produced and stored is unused by the organizations. Another being the quality of big data, many researchers today question the efficacy and quality of massive amounts of data being collected each day. Some argue that the efficiency of current approaches to detect anomalies is being reduced because of the huge volume of data being accumulated [5]. Others [6] highlight the importance of data quality assessment because of the noise. Another study [7] says the goal of data cleansing is to ensure quality data analytics. Thus, it becomes important to understand the percentage of quality data, and/or data being used out of all the data being generated and collected.

Another problem that arises in businesses with the collection of big data is data hoarding [8][9][10][11]. Organizations fear missing out on potential future analytical opportunities, thus 50-90% [8] of stored data offering no value to business at the time of collection. There are numerous drawbacks associated with data hoarding including increased operational inefficiencies,

elevated security risks, and a considerable increase in environmental costs due to the energy-intensive nature of data centers. Thus, the question: Does big data really need to be big?

The goal would be to achieve the answer to these questions by developing a framework to assess big data quality. One of the factors that influences the quality of the data is the business domain selected to collect the data from [6]. Thus, after selecting a business domain, the next step is to experiment with and establish a set of criteria for evaluating and enhancing big data quality in business contexts. Finally, the usefulness of the whole dataset can be determined when comparing the values of the data (Big Data in this case). The initial hypothesis is that this can be achieved by using Apache Spark, a popular framework for working with big data.

For the purposes for demonstration of the proposed Dark Data Transformation Module (DDTM) synthetic data is being used here (Modern Healthcare and Research Data generated by Synthea™).

II. LITERATURE REVIEW

The term called dark data [12] appeared several times while conducting literature review, this is data collected by businesses for their regular operations but remains widely unused for analytics or decision-making processes [13]. Just like big data, dark data can be structured, unstructured and semi-structured and can come from sources such as transaction histories, images, videos, text files, logs and more [14]. Some early studies categorize dark data into three types [15]. First, the data that is not being collected by anyone. Second, data that is collected, but proves difficult to access at the right time. And third, data that is collected but has not yet been productized.

One of the early representations of dark data is shown in figure 1 [15] which also shows the potential that lies beneath the surface of enterprise data.

According to some dark data is a subset of big data that remains stored in the data warehouses of companies. Moreover, some estimates say 43-50% [16][17] of the data collected goes into dark state while others say that at least 90% of data in companies can be considered dark [18]. Tech giants such as Microsoft also question if big data is turning into dark data. Others estimate that only 15% of the overall data is business critical data and the rest is either redundant, obsolete and trivial (ROT) or dark [13].

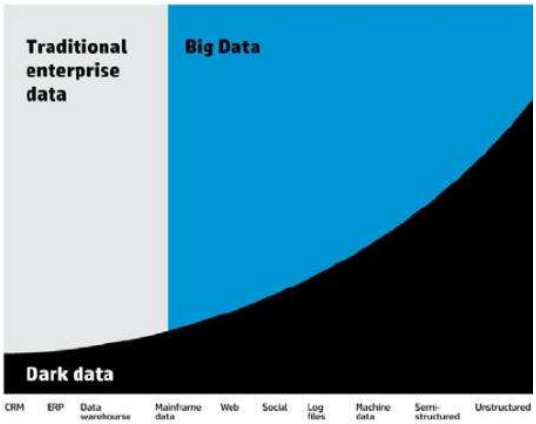


Figure 1: Percentage of dark in Big Data and Traditional Enterprise data

One of the newer representations of dark data is shown in figure 2 [19].

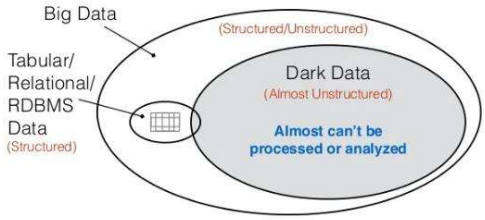


Figure 2: Composition of dark data

Research shows a few reasons for this phenomenon as described below,

1) Distributed storage systems make it difficult for teams to understand the actual amount of data present in the data lakes [14] [18]
2) There exists a lack of awareness when it comes to existence, or the potential of data being collected [14]
3) Many companies hoard data without thinking about future or analytics usage because it is easy and cheap to store data [18]
4) Some businesses store data to fulfil compliance policies that mandate organizations to store sensitive information up to a certain period [14] [18]
5) Another issue is the relevance of data. Some argue that data collected is rendered irrelevant to business use cases after some time due to changing needs and demands [18]

Not only is data hoarding inefficient, but it also comes with its own risks. It can be a very easy target for hackers to exploit sensitive information. In addition to the security risks, there are a few indirect costs associated with storing unused data for longer periods of time. This includes the opportunity costs of storing massive data. By not utilizing this data, businesses also lose out on opportunities to drive revenue and innovation. Dark data can be employed to build machine learning models, which in turn could predict better outcomes or help companies serve their customers better [14] [20].

After careful analysis of the studies reviewed thus far, an answer to the first question about the amount of data needed for analytics can be drafted. Although a rough estimate, the belief remains that only 10-20% of the data is used for analytics.

To address the second question, it cannot be guaranteed that the data being collected and stored is entirely of no use until the potential of dark data is tested. As researchers suggest dark data might lead us to better machine learning and deep learning models leading to more accurate analysis and predictions. In fact, technicians [21] have already started to explore the opportunities of using AI for dark data identification and increasing data privacy. They believe that most of this dark data contains information that is sensitive in nature and can lead to security breaches if not identified and secured at early stages.

After identification of dark data, if better analytic results can be achieved by feeding this data to machine learning or deep learning models, the answer to the third question would be ‘no’ since increasing the dataset could potentially prove beneficial. Given the nature or size of dark data, it may be viable to feed the dataset to a large language model (LLM),

but this remains an exploratory field. Moreover, the difficulty of getting the data verified by domain experts before training an LLM with it would be a necessary step.

One of the challenges in finding answers to key questions is the discovery of dark data. From the definition of dark data itself it is revealed that it is a form of data that businesses do not know exists, implying that dark datasets might not be available publicly for testing and experimenting. This also makes dark data business specific, thus, it becomes important to consider this when proposing a model or experiment.

Another interesting question that can be posed at this stage in addition to the existing three questions is: *How* can value be extracted from dark data?

III. PROPOSED MODEL

Over recent years, unutilized data (dark data) has gained the attention of researchers as well as businesses. Some researchers emphasize the governance of dark data to reduce security breaches [22], while others suggest management techniques by understanding the characteristics of dark data specific to the business they are dealing with [23]. Another study suggests that businesses should leverage dark data as an opportunity rather than seeing it as a burden [24]. There have been recorded instances where the sheer volume of dark data is so large (even for smaller enterprises) [26] that the cost of traversing through the data can be a huge hike in expenditure.

The Dark Data Management Repercussion Model [26] investigates both the positive and the negative results of handling dark data. As seen in figure 3 [26] *lighting*, the suppression of dark data is one method of handling it, involves utilizing unused data, mitigating threats and ensuring competitiveness.

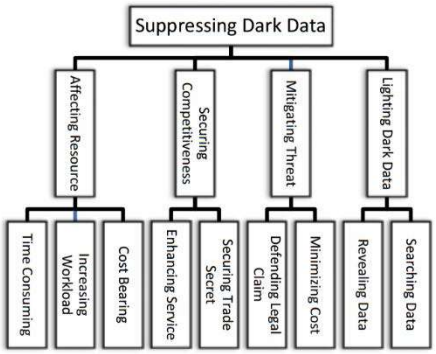


Figure 3: Suppression of dark data

Redundant, noisy, and unstructured data is often classified as dark data, yet analyses suggest it can boost business revenue [26]. The FAIR standard (Findable, Accessible, Interoperable, Reusable) is a common metric for evaluating data [27], but usability (U) is considered a better measure for dark data, as truly usable data would be categorized as gray rather than black [26].

Enhancing dark data discoverability relies on rich metadata (MD), which includes technical attributes like filesystem size, descriptive details such as keywords, titles, and creators, process-related information like computing environment and software used, and domain-specific descriptions covering aspects like simulated systems, spatial resolution, and controlled variables [28].

While these methods may improve the visibility of black data, they are not sufficient to eliminate its 'dark' classification. Building upon existing frameworks, the Quality-Driven Data Reduction (QDDR) model is proposed to address key questions regarding data volume, utility, and analytics efficiency.

To develop a data quality assessment method, Data Quality Dimensions [28] need to be understood. This is a set of data quality attributes that represent a single aspect or construct of data quality. The most common data quality dimensions (DQD's) are Accuracy, Completeness, Consistency, Freshness, Uniqueness, and Validity. Another related concept is Data Reflectivity [29], which indicates how data mirrors reality.

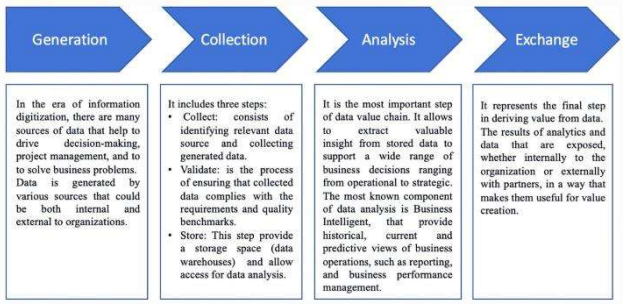


Figure 4: Data lifecycle: from creation to sharing insights

Regarding the transformation of raw data into valuable insights that accurately reflect reality, Faroukhi et al. [30] proposed a data value chain, as shown in Figure 4. This highlights the critical role of analysis in assigning "value" to data. If comparable analysis results can be achieved, the value of data remains intact.

Significant efforts have been made in assessing data quality and detecting anomalies. Yusufi et al. [31] introduced a Data Value Assessment (DVA) process to define and quantify data value in production. During this process, data metrics were rated manually—an efficient approach for optimizing the DVA process when handling user data. However, in the absence of collaboration with domain experts, an automated mechanism [32][33] is preferred.

A review of relevant work revealed that machine learning and deep learning models are widely applied for data value and quality assessments [33][34][35][36], demonstrating effectiveness even on unstructured data [37]. Other approaches focus on anomaly detection [38], feature selection [39], and blockchain-based solutions [40].

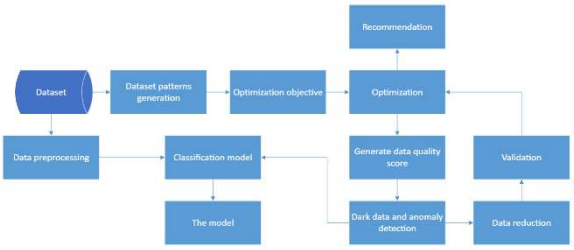


Figure 5: Dark Data optimization pipeline

The conceptual architecture for data quality enhancement is

presented in Figure 5. The first step involves generating patterns from datasets, which can be achieved by calculating statistical measures such as variance, entropy, and correlation. These factors are then used to formulate an optimization objective, such as a linear objective function:

$$w_1 a_1 + \cdots + w_n b_n = Quality\ Score$$

Based on the primitive function, a Quality Score is generated for dataset units (e.g., each column), where the lowest-scoring units are identified as data quality anomalies. These data units are reduced to form a new dataset for validation. Validation involves performing a specific data analysis task (e.g., classification of cyber-attacks). If the analytical results remain consistent with the original dataset, the detection method is deemed effective.

The function is then optimized to determine the values of w (weights) that maximize the correlation between the Quality Score and the performance of analysis tasks. Meanwhile, the reduced dataset can be utilized to train a binary classification model. By implementing this process, a recommendation is proposed based on the optimized equation that defines the significance of each dataset pattern, along with a trained classification model for anomaly detection. Both outputs contribute to assessing data quality and leveraging dark data.

IV. ANALYSIS

To find the potential in dark data, it is important to first identify and collect it. One interpretation of dark data is that it consists of information companies are unaware of, making it challenging to identify available dark datasets online. Moreover, identification of dark data might even be a bigger challenge than using it for analysis. As per studies [16][17][18], companies are not aware of the existence of dark data in their data lakes. Thus, this paper tries to implement a pipeline that converts data received in different formats into a single format. After identifying the relationships between different objects, data engineers can combine the data and/or use this data to identify patterns and trends.

In the model developed, a dataset is first identified that could be used to replicate dark data (unused data found in data lakes). The goal is to expose this untapped data for the use of the company. The one used here is a synthetic version of a patient's medical records. The data picked is then converted into several data formats to simulate how unused data may be found in a company's data warehouse. The next step is to collect the different formats of data allocated from different origins within the company. This data is then processed using the Dark Data Transformation Module (DDTM), where three types of data formats - .txt, .json, and .h5 - are handled.

The .txt goes line by line and collects the value separated by their delimiters, then it sorts them into rows in a csv file. In the .json converter, the data is first converted into a raw JSON message. The header names are then separated, and a CSV file is created with all the header names. The values from the JSON are placed into the rows of the CSV, and every field is saved as a string in the CSV. For the h5 files, where the data is hierarchical, it is important to understand the structure of the file. This process requires manual intervention to find the level of the hierarchy (called block in this case) that the data resides in and then a systematic extraction of the blocks. Once the level has been decided, it is a simple iteration over blocks of lists separated based on the data type of each field.

The source files are first inspected manually to understand configurations such as delimiter used, end of line – carriage return or line feed or a combination of both and quoted text or unquoted text. Manual inspection is important in order to automate this task using a tool such as Azure Data Factory (ADF), that requires this metadata, where an Extract-Transform-Load (ETL) pipeline can be built to execute different scripts to handle different data formats.

After the conversion, it is important to understand the relationship between different objects or entities. This can be achieved by creating an entity-relationship diagram for labeled dataset (see Appendix) or using a clustering algorithm on an unlabeled dataset. After this, QA validation can be performed against the dataset being used on a regular basis.

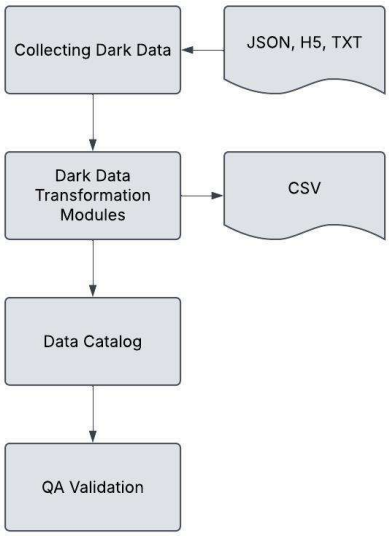


Figure 6: Flow of our proposed model

The result achieved so far is different than expected in terms of unlocking the potential of dark data. The expectation was to be able to find a dark dataset and test it with a machine learning or a deep learning model for an analysis task, however finding dark dataset proved to be difficult. Thus, synthetic data had to be used. However, it did solve another interesting problem of finding dark data and giving it more visibility from an organization perspective. Having a pipeline dedicated to transforming dark data into a usable format could be a good start for companies to start exploring the potential. Once it has been established, the data could then be utilized for further analysis.

Thinking from an organization's perspective, converting unused data to something reusable could be very crucial and could add a lot of value to the business. At the same time, performing conversion on unpredictable or untapped data could be a cost extensive task because of the huge volume of data accumulated over time. The execution time of the conversion might also be so huge that by the time it finishes, the data is rendered useless. Moreover, on a big scale, the solution might not be generalizable for multiple objects due to increased complexity.

Some organizations might also question the worth of going through this trouble for something that they may not even find. Dark data inherently does not guarantee better results. It is by

experimentation and validation that one could predict its usability for analytical purposes.

V. CHALLENGES & LIMITATIONS

One of the challenges is the general adaptation of using dark data. Businesses may not want to invest resources in something that is not guaranteed. The cost of performing tasks such as collection, conversion, cataloging and validation might be steeper than storing the data in the enterprise data lakes abundantly. Another challenge is that businesses might not be aware of the presence of sensitive or personally identifiable data within their data lakes and might accidentally end up exposing unwanted information. This could potentially lead to governance issues.

One of the other challenges that could be faced by Data Engineers while working with dark data is dealing with highly unstructured data or complexly structured data. For example, multiple objects within an object in a JSON file would require a thorough examination to understand the relationships between different attributes as well as to write the module for conversion. This would again require a huge amount of time to be invested by organizations at a huge cost.

Limitations of the process are listed below,

1. While the Synthetic Patient Data provides a virtual domain for experimentation, the current approach requires a more stimulated data collection method to mirror real-world scenarios; additionally, transforming only 16 files is significantly less representative than actual data volumes.
2. Although CSV, H5, JSON, and TXT are popular file formats across business domains, there remains substantial potential for discovering and incorporating additional data formats. Dark data can potentially exist in numerous storage forms, yet this project exclusively focused on text and numeric data types.
3. The Synthetic Patient Data offers contextually related content, which differs markedly from real-world organizational data that frequently contains disparate, potentially irrelevant information requiring comprehensive manual inspection.

VI. CONCLUSION & FUTURE WORK

This research goes in depth to explore critical challenges faced when managing dark data, in effort to question the conventional perception that big data must always be voluminous. In the proposed model, the Quality-Driven Data Reduction (QDDR) framework was presented which is a new approach to identifying, transforming, and potentially leveraging seemingly unused organizational data.

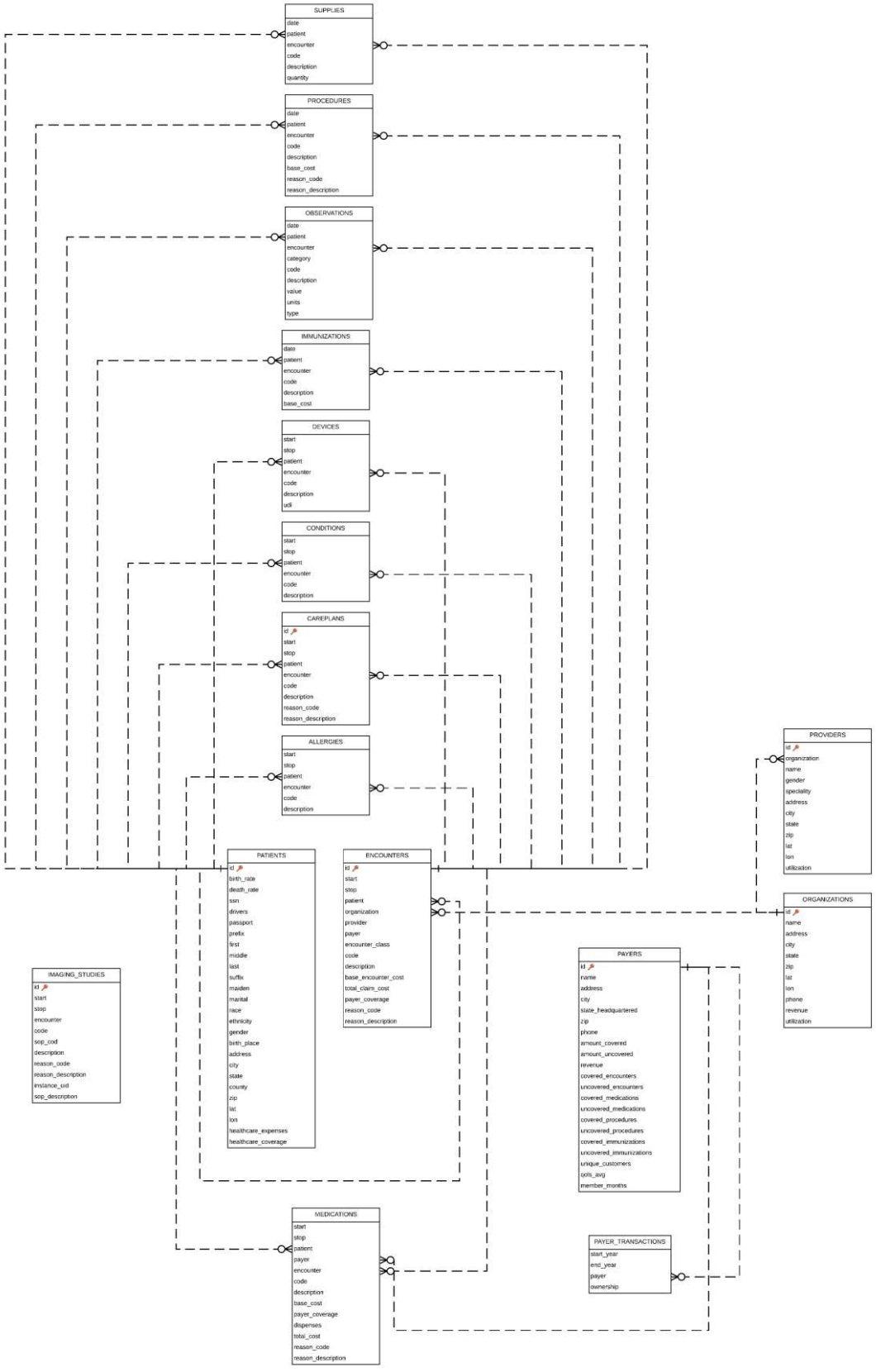
The untapped potential of dark data became apparent during research which revealed that organizations potentially store 50-90% of data in a "dark" state. The Dark Data Transformation Module (DDTM) developed will offer an initial pipeline to convert disparate data formats into a unified, analyzable structure. The synthetic dataset used here provides proof of concept, and the research done emphasizes the complexity of handling dark data appropriately.

Finally, dark data's value in big data is not guaranteed but it requires systematic exploration, validation, and careful assessment using methods discussed in this paper.

In Big Data, dark data shows significant promise, possible *future* avenues include:

1. Machine Learning Integration – Develop sophisticated machine learning models which are able to assess and extract value from dark dataset, merging both the models introduced in this paper.
2. Real World Domain Validation – This paper uses synthetic datasets to simulate and handle dark data, to confidently validate the generalizability of the models/ frameworks developed, multiple experiments need to be conducted using real world organizational data.
3. Sensitive Data Handling – As mentioned earlier, there is often a significant concern for sensitive or personally identifiable information found in dark data, creating robust mechanisms for automatically identifying this information would help counter this issue.
4. Large Language Model Exploration – Explore the potential of training large language models (LLMs) using dark data, which in turn can be used to develop rigorous verification protocols to ensure data quality and reliability.
5. Cost-Benefit Analysis Framework – For organizations and companies a comprehensive framework to evaluate the economic feasibility of dark data transformation, considering the cost of lightening the data against the probability of the usable data.

These are some of many research directions that can be explored to move closer to either transforming dark data into a strategic organizational asset or eliminating it to improve quality of data collected.


SUPPLIES
date
patient
encounter
code
description
quantity
PROCEDURES
date
patient
encounter
code
description
base_cost
reason_code
reason_description
OBSERVATIONS
date
patient
encounter
category
code
description
value
units
type
IMMUNIZATIONS
date
patient
encounter
code
description
base_cost
DEVICES
start
stop
patient
encounter
code
description
udi
CONDITIONS
start
stop
patient
encounter
code
description
CAREPLANS
id
start
stop
patient
encounter
code
description
reason_code
reason_description
ALLERGIES
start
stop
patient
encounter
code
description
PROVIDERS
id
organization
name
gender
speciality
address
city
state
zip
lat
lon
utilization
PATIENTS
id
birth_rate
death_rate
ssn
drivers
passport
prefix
first
middle
last
suffix
maiden
marital
race
ethnicity
gender
birth_place
address
city
state
county
zip
lat
lon
healthcare_expenses
healthcare_coverage
ENCOUNTERS
id
start
stop
patient
organization
provider
payer
encounter_class
code
description
base_encounter_cost
total_claim_cost
payer_coverage
reason_code
reason_description
ORGANIZATIONS
id
name
address
city
state
zip
lat
lon
phone
revenue
utilization
IMAGING_STUDIES
id
start
stop
encounter
code
sop_cod
description
reason_code
reason_description
instance_uid
sop_description
PAYERS
id
name
address
city
state_headquartered
zip
phone
amount_covered
amount_uncovered
revenue
covered_encounters
uncovered_encounters
covered_medications
uncovered_medications
covered_procedures
uncovered_procedures
covered_immunizations
uncovered_immunizations
unique_customers
qols_avg
member_months
MEDICATIONS
start
stop
patient
payer
encounter
code
description
base_cost
payer_coverage
dispenses
total_cost
reason_code
reason_description
PAYER_TRANSACTIONS
start_year
end_year
payer
ownership

References

1. H. E. Pence, "What is Big Data and Why is it Important?," Journal of Educational Technology Systems, vol. 43, no. 2, pp. 159-171, 2014.
2. S. Khan, "COMP-8157 lecture 2," School of Computer Science, University of Windsor, Windsor, ON, Canada, 2025.
3. "What is Big Data," Google Cloud. [Online]. Available: https://cloud.google.com/learn/what-is-big-data [Accessed: Feb. 23, 2025].
4. R. J. T. Morris and B. J. Truskowski, "The evolution of storage systems," IBM Systems Journal, vol. 42, no. 2, pp. 205-217, 2003.
5. R. A. A. Habeeb, F. Nasaruddin, A. Gani, I. A. T. Hashem, E. Ahmed, and M. Imran, "Real-time big data processing for anomaly detection: A survey," International Journal of Information Management, vol. 45, pp. 289-307, 2019.
6. S. Sarker, M. S. Arefin, M. Kowsher, T. Bhuiyan, P. K. Dhar, and O.-J. Kwon, "A comprehensive review on big data for industries: challenges and opportunities," IEEE Access, vol. 11, pp. 744-769, 2022.
7. A. M. Rahmani, E. Azhir, S. Ali, M. Mohammadi, O. H. Ahmed, M. Y. Ghafour, S. H. Ahmed, and M. Hosseinzadeh, "Artificial intelligence approaches and mechanisms for big data analytics: a systematic study," PeerJ Computer Science, vol. 7, p. e488, 2021.
8. M. Hawker, "A CDO call to action: Stop hoarding data—save the planet," Forbes Tech Council. [Online]. Available: https://www.forbes.com/councils/forbestechcouncil/2023/07/20/a-cdo-call-to-action-stop-hoarding-data-save-the-planet/ [Accessed: Jan. 28, 2025].
9. D. B. Laney, "The data purge: An era of defensible retention and data minimization," Forbes. [Online]. Available: https://www.forbes.com/sites/douglaslaney/2023/06/13/the-data-purge-an-era-of-defensible-retention-and-data-minimization/ [Accessed: Jan. 28, 2025].
10. N. Eide, "Leftover data lurks across the enterprise, creating a business risk," Cybersecurity Dive. [Online]. Available: https://www.cybersecuritydive.com/news/data-retention-cyber-risk/647131/ [Accessed: Feb. 23, 2025].
11. S. Lynn, "Halt data hoarding in 2025 with these tips from an IT exec," MES Computing. [Online]. Available: https://www.mescomputing.com/news/business/halt-data-hoarding-in-2025-with-these-tips-from-an-it-exec [Accessed: Jan. 28, 2025].
12. "Dark Data: Discovery, Uses & Benefits of Hidden Data," Splunk. [Online]. Available: https://www.splunk.com/en_us/blog/learn/dark-data.html [Accessed: Feb. 23, 2025].
13. "What is dark data? Uncovering vulnerable data," BigID. [Online]. Available: https://bigid.com/blog/what-is-dark-data/ [Accessed: Feb. 23, 2025].
14. "What is dark data?," IBM Think. [Online]. Available: https://www.ibm.com/think/topics/dark-data [Accessed: Feb. 23, 2025].
15. L. Akbar, K. al-mutahr, and M. Nazeh, "Aligning IS/IT with business allows organizations to utilize dark data," Eurasian Journal of Analytical Chemistry, vol. 13, pp. 137-145, 2018. [Online]. Available: https://www.researchgate.net/publication/332138489_Aligning_ISIT_with_business_allows_organizations_to_utilize_dark_data [Accessed: Feb. 23, 2025].
16. "Future of big data: Insights & trends 2024," ForageAI, LinkedIn. [Online]. Available: https://www.linkedin.com/pulse/future-big-data-insights-trends-2024-forageai-zahwc [Accessed: Feb. 23, 2025].
17. "The state of dark data: Industry leaders reveal the gap between AI's potential and today's data reality," Splunk. [Online]. Available: https://www.splunk.com/en_us/form/the-state-of-dark-data.html [Accessed: Feb. 23, 2025].
18. A. S. George, V. Sujatha, A. S. H. George, and T. Baskar, "Bringing Light to Dark Data: A Framework for Unlocking Hidden Business Value," Partners Universal International Innovation Journal (PUIIJ), vol. 1, no. 4, pp. 1-40, Jul.-Aug. 2023. DOI: 10.5281/zenodo.8262384.
19. "Big data vs. dark data," Datadition. [Online]. Available: https://datadition.com/big-data-vs-dark-data/ [Accessed: Feb. 23, 2025].
20. E. Charran, "Armchair Architects: Is Big data turning into dark data?," Microsoft Tech Community Blog. [Online]. Available: https://techcommunity.microsoft.com/blog/azurearchitectureblog/armchair-architects-is-big-data-turning-into-dark-data/3725726 [Accessed: Feb. 23, 2025].
21. M. Nozari and U. Majumder, "Mastering the dark data challenge: Harnessing AI for enhanced data governance and quality," Enterprise Knowledge. [Online]. Available: https://enterprise-knowledge.com/mastering-the-dark-data-challenge-harnessing-ai-for-enhanced-data-governance-and-quality/ [Accessed: Feb. 23, 2025].
22. P. Dimitrov and D. Chikalanov, "The role of governance in managing dark data to reduce security risks," in *Unlocking Hidden Value: A Framework for Transforming Dark Data in Organizational Decision-Making*, A. Leogrande, Ed., LUM University Giuseppe Degennaro, Casamassima, Bari, Puglia, Italy, pp. 3-4, 2016.
23. A. F. M. Ajis, J. M. Jali, I. Ishak, and Q. N. Harun, "Dark data management framework: A grounded theory approach," in *Unlocking Hidden Value: A Framework for Transforming Dark Data in Organizational Decision-Making*, A. Leogrande, Ed., LUM University Giuseppe Degennaro, Casamassima, Bari, Italy, pp. 3-4, 2022.
24. R. Corallo, A. Lazoi, and G. Secundo, "Dark data management in the manufacturing industry: A framework for optimizing industrial processes," in *Unlocking Hidden Value: A Framework for Transforming Dark Data in Organizational Decision-Making*, A. Leogrande, Ed., LUM University Giuseppe Degennaro, Casamassima, Bari, Italy, pp. 15-20, 2021.
25. A. F. M. Ajis, J. M. Jali, I. Ishak, and Q. N. Harun, "Enlightening the repercussion of dark data management towards Malaysian SMEs sustainability," *SSRN Electronic Journal*, pp. 1-10, 2022. [Online]. Available: https://www.researchgate.net/publication/374897395_Enlightening_the_Repercussion_of_Dark_Data_Management_towards_Malaysian_SMEs_Sustainability [Accessed: Feb. 25, 2025].
26. A. Almeida, M. Torres-Espin, A. J. Ferguson, and M. G. Fehlings, "Excavating FAIR data: The case of the multicenter animal spinal cord injury study (MASCIS), blood pressure, and neuro-recovery," *Neuroinformatics*, vol. 19, no. 1, pp. 117-130, Mar. 2021, doi: 10.1007/s12021-021-09520-8.
27. B. Schembera, "The dark side of data management," Apr. 10, 2019. [Online]. Available: https://www.researchgate.net/publication/355545901_The_Dark_Side_of_Data_Management. [Accessed: Feb. 25, 2025].
28. A. Ramasamy, S. Chowdhury. "Big data quality dimensions: a systematic literature review," in JISTEM-Journal of Information Systems and Technology Management, vol. 17, pp. e202017003, 2020.
29. C. Strzelecka. "Critical data studies meets discard studies: Waste data reflectivity in digital urban waste tracking system," in Convergence, vol. 30, no. 6, pp. 2109–2130, 2024.
30. Faroukhi, A., et al. "Big data monetization throughout Big Data Value Chain: a comprehensive review," in Journal of Big Data, vol. 7, pp. 1–22, 2020.
31. Yusufi, Z., et al, "Data Value Assessment in Semiconductor Production: An Empirical Study to Define and Quantify the Value of Data," in Proceedings of the 6th International Conference on E-Commerce, E-Business and E-Government, 2022, pp. 109–116.
32. Li, N., et al, "Ocean data quality assessment through outlier detection-enhanced active learning," in 2023 IEEE International Conference on Big Data (BigData), 2023, pp. 102–107.
33. G. Mylavarapu, J. Thomas, K. Viswanathan, "An Automated Big Data Accuracy Assessment Tool," in 2019 IEEE 4th International Conference on Big Data Analytics (ICBDA), 2019, pp. 193-197.
34. Shinkuma, R., et al. "Data assessment and prioritization in mobile networks for real-time prediction of spatial information using machine learning," in EURASIP Journal on Wireless Communications and Networking, vol. 2020, no. 1, pp. 92, 2020.
35. Sarker, I., et al. "Cybersecurity data science: an overview from machine learning perspective," in *Journal of Big data*, vol. 7, pp. 1–29, 2020.
36. S. Juddoo, C. George, "A Qualitative Assessment of Machine Learning Support for Detecting Data Completeness and Accuracy Issues to Improve Data Analytics in Big Data for the Healthcare Industry," in 2020 3rd International Conference on Emerging Trends in Electrical, Electronic and Communications Engineering (ELECOM), 2020, pp. 58-

66.

37. Kaliyaperumal, G., et al, "Adaptive Framework for Comprehensive Quality Assessment in Unstructured Big Data," in 2024 7th International Conference on Contemporary Computing and Informatics (IC3I), 2024, pp. 1522-1528.
38. E. Widad, E. Saida, Y. Gahi. "Quality Anomaly Detection Using Predictive Techniques: An Extensive Big Data Quality Framework for Reliable Data Analysis," in IEEE Access, vol. 11, pp. 103306-103318, 2023.
39. A. Nayak, B. Božić, L. Longo. "Data Quality Assessment and Recommendation of Feature Selection Algorithms: An Ontological Approach," in Journal of Web Engineering, vol. 22, no. 1, pp. 175-196, 2023.
40. Y. Wu, K. Sha, K. Yue, "Poster: Blockchain-Enabled Federated Edge Learning for Big Data Quality Assessment," in 2022 IEEE/ACM 7th Symposium on Edge Computing (SEC), 2022, pp. 284-285.