DOI: 10.15439/2024F0005
ISSN 2300-5963 ACSIS, Vol. 40

QEDrants – Data Quality Quadrants for Business Users and Decision-Makers

Alina Powała*0009−0009−0268−3582, Dominik Ślęzak*†0000−0003−2453−4974

*QED Software, Mazowiecka 11/49, 00-052 Warsaw, Poland
Email: {alina.powala,dominik.slezak}@qedsoftware.com
†Institute of Informatics, University of Warsaw, Banacha 2, 02-097 Warsaw, Poland
Email: slezak@mimuw.edu.pl

***Abstract*—The adoption of artificial intelligence (AI) in business is often hindered by the complexity of data quality assessment. This paper introduces the quadrant-based data quality representation framework, which evaluates data assets based on two complementary dimensions: Data Integrity (accuracy and reliability, akin to Gartner's "Ability to Execute") and Data Coverage (breadth and comprehensiveness, similar to "Completeness of Vision"). The framework categorizes data into four groups: *Pure Gold* (AI-ready), *Sleeping Giants* (high integrity, low coverage), *Unpolished Diamonds* (high coverage, low integrity), and *Hitchhikers* (low integrity, low coverage). Each such quadrant provides actionable insights for business users, helping them prioritize data assets for AI readiness, identify data cleaning tasks, balancing costs and value realization by focusing on the right data. Given the roots of this idea in QED Software's technology experiences, we call the proposed quadrants as *QEDrants*.**

***Index Terms*—Data Quality, Data Integrity, Data Coverage, AI Readiness, Data Management, Decision Support, Cost Optimization**

I. Introduction

CAN businesses effectively manage data without constant reliance on data scientists? The purpose of this work is not to diminish the critical role of data scientists but to address the persistent gap between business stakeholders and technical teams. Even in organizations with strong artificial intelligence (AI) capabilities, this disconnect often leads to inefficiencies and misaligned goals. Bridging this gap requires establishing a common ground where business users can better understand technical concepts, and data scientists can align their work more closely with business objectives [1].

In particular, many organizations encounter the challenge of ensuring the quality of data, which is crucial for producing impactful AI outcomes. For non-technical stakeholders, assessing data quality is complex, making it hard to determine when data is ready to support business-critical AI applications. A part of the domain of AI refers to machine learning (ML)[1], wherein the challenges of poor data quality are especially well-understood. However, this problem is broader and does not refer only to the methods that we would call pure ML [2].

To address the above gap from the data quality perspective, we introduce the framework called *QEDrants*[2], which categorizes data assets into four groups based on two business-friendly dimensions: *Data Integrity* and *Data Coverage*.

Data Integrity corresponds to data accuracy and reliability. It assesses whether the data adheres to predefined rules and standards, such as valid formats (e.g., properly structured dates) and logical consistency (e.g., non-negative values in the age fields). High integrity ensures the data is free from errors and can be trusted for analysis and decision-making. Data Coverage, on the other hand, measures the completeness and comprehensiveness of the dataset. High coverage indicates that the dataset captures the full scope of the domain it describes, ensuring no critical information is missing. These two metrics are not opposing forces but rather complementary drivers of data quality. Together, they determine the data utility in delivering actionable insights and value. Just as both execution and vision are crucial for a business to thrive, integrity and coverage are essential for data to achieve its full potential.

We define QEDrants as follows:

- *Pure Gold*: High integrity and high coverage data, ideal for direct application in AI models.
- *Unpolished Diamonds*: High coverage but lower integrity data, representing assets that are rich in content but may need refinement for reliable AI use.
- *Sleeping Giants*: High integrity but low coverage data, indicating well-curated yet incomplete data sources that could benefit AI if augmented.
- *Hitchhikers*: Low integrity and low coverage data, representing low-priority assets that are generally unsuitable for AI applications.

The goal of QEDrants is to deliver the quadrant-style visualization with a clear, actionable view of data quality, highlighting areas where data can best serve AI objectives and where it requires improvement. This way, QEDrants can provide business users with a practical tool to prioritize data curation efforts, focusing on areas where investment will yield the greatest gains in AI readiness. The goal of QEDrants is also to emphasize the *value realization potential* of data, demon-

[1] Although AI and ML can be considered as two separate domains, in this paper – for simplicity – we use the acronym "AI" to cover both of them.

[2] Name inspired by Gartner's *Magic Quadrants*TM https://www.gartner.com/.

©2024

strating how data assets contribute to usability, actionability, and value extraction – concepts aligned with the "5 V's of Big Data" (volume, velocity, value, variety, veracity) [3].

The remainder of this paper is organized as follows. Section II discusses related work on data quality in AI, barriers in AI adoption, and existing data quality frameworks. Section III recalls broader inspirations and connections, including insights from Gartner's frameworks, the concept of Total Cost of Ownership (TCO) for data processing systems, and related disciplines such as data governance and data security. Section IV introduces the conceptual and architectural background for the methodology used to define QEDrants. Section V presents case studies illustrating the application of our framework in various business contexts. Finally, Section VI concludes the paper.

II. Related Work

Data quality has been a long-standing area of research within AI, as the success of AI systems is closely tied to the validity of learning. In recent years, a substantial body of work has addressed the critical aspects of data quality in AI, identifying key challenges and proposing frameworks for evaluating data quality across different domains. However, while numerous approaches to data quality exist, many are tailored primarily for data scientists, leaving business stakeholders with limited accessibility to those methodologies.

A. Data Quality in AI

Research on data quality in AI emphasizes its pivotal role in ensuring reliable and unbiased outputs. Early work focused on core data quality dimensions such as accuracy, completeness, consistency, timeliness, and relevance [4], [5]. These dimensions remain foundational for assessing data quality in modern AI applications, as they directly affect model performance, interpretability, and the capacity to generalize. Several studies link data quality issues to AI model training and deployment challenges. Poor data quality can lead to model overfitting and inaccuracies, ultimately diminishing the value of AI insights for business stakeholders [6].

B. Barriers to AI Adoption

Despite the advancements in data quality research, barriers to AI adoption in business persist. They are attributed to the lack of accessible and interpretable data preparation and assessment frameworks. Non-technical users, including business managers and decision-makers, often lack the tools that are needed to assess data quality or improve data readiness for AI applications. Studies indicate that without straightforward methods for evaluating data, organizations face increased costs, prolonged implementation timelines, and potential failures in AI deployment due to suboptimal data preparation [7]. Moreover, traditional data quality frameworks typically emphasize technical dimensions without considering usability in business. This technical focus can lead to an "AI unreadiness," where data quality needed for effective AI outcomes is not in place, resulting in limited confidence in AI systems.

C. Existing Frameworks

Several frameworks have been developed to provide a systematic approach to data quality evaluation, including Total Data Quality Management (TDQM) [8], Data Quality Assessment (DQA) [9], and others based on international standards like ISO/IEC 25012. These frameworks define comprehensive methodologies for assessing and improving data quality across dimensions (see [5] for a survey of approaches). However, they are geared towards data engineers and scientists, involving complex metrics and extensive data profiling procedures that may be cumbersome for business users. Furthermore, some recent frameworks include domain-specific quality models for healthcare, finance, and retail [10]. While these models add valuable insights into data quality needs for AI, they still tend to require high technical proficiency and they do not address the accessibility requirements of non-technical users.

D. Gaps in Business-Friendly Evaluation

The existing literature on data quality frameworks reveals a clear gap in models that are accessible to non-technical users and aligned with their business goals. As we have already discussed, this is part of a broader issue of the gap between business and AI specialists, which still exists even in the case of relatively large and mature companies. Traditional frameworks focus on rigorous technical assessment, which is essential in data science but lacks usability for stakeholders who lack deep technical knowledge. As a result, many organizations face challenges in bridging the gap between data engineering teams and business decision-makers.

To support AI adoption in business, there is a need for simplified, business-oriented frameworks that can help non-technical users understand and prioritize data quality issues [11]. Such a framework would empower business leaders to make informed decisions about data readiness for AI, minimizing technical barriers and accelerating AI adoption. QEDrants will address this gap as they are designed to be accessible to business stakeholders, facilitating the identification of data quality issues with minimal technical complexity.

III. Connections to Other Domains

The previous section focused on related work concerning the importance and measurement of data quality. This part expands the scope to explore broader inspirations. Although the areas considered below are not directly tied to data quality, they intersect in meaningful ways, influencing and shaping one another within the data management ecosystem.

A. Gartner's Magic Quadrants

Gartner is widely recognized for its proprietary methodologies, including the concept of Magic Quadrant$^{\text{TM}}$ which is famous primarily because it provides a clear, visual framework for comparing technology providers in various industries, simplifying the decision-making process for businesses (see e.g. Fig. 1). It breaks down complex market analyses into a simple, two-axis chart, categorizing vendors into four types –

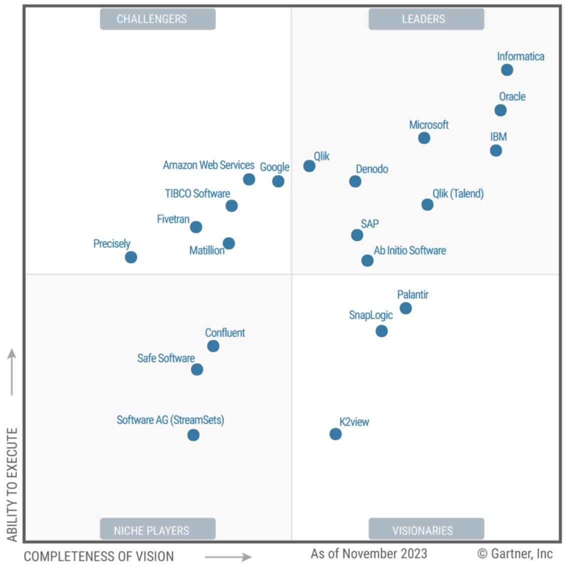

Fig. 1: Gartner's Magic Quadrant^{TM} for data integration tools (https://www.informatica.com/content/dam/informatica-com/en/image/misc/data-integration-magic-quadrant-2023.png)

Leaders, Challengers, Visionaries, and Niche Players. The vertical axis, "Ability to Execute," evaluates a vendor's capacity to deliver on its promises, including product quality, customer support, and financial performance [12]. The horizontal axis, "Completeness of Vision," assesses a vendor's understanding of current and future market dynamics, innovation, and alignment with customer needs. This clarity makes it easier for companies to determine the competitive landscape at a glance.

We want QEDrants to leverage a two-axis model too. In our case, the focus is on Data Integrity and Data Coverage – two forces that work together to assess the data readiness for AI applications. Unlike Gartner's model, which primarily evaluates vendor performance, QEDrants apply these dimensions to data quality, offering a unique perspective on how organizations can use and improve their data to support AI initiatives. In this context, "Ability to Execute" from the Magic Quadrant^{TM} framework aligns conceptually with Data Integrity. Just as "Ability to Execute" reflects a vendor's capacity to deliver on promises, Data Integrity measures the reliability and accuracy of the data, ensuring it is fit for purpose. Furthermore, "Completeness of Vision" corresponds to Data Coverage. In the Magic Quadrant^{TM} model, "Completeness of Vision" means a vendor's forward-looking strategy and understanding of market trends. In QEDrants, Data Coverage assesses the comprehensiveness and representativeness of the data, ensuring it captures all necessary dimensions for effective AI deployment.

These two dimensions – Data Integrity and Data Coverage – are not opposites but rather complementary forces. Together, they provide a holistic view of data quality, ensuring organizations can trust their data and rely on its breadth. To our best knowledge, no Gartner-inspired quadrant visualization has been applied specifically to data quality assessment. While Gartner has utilized similar visual frameworks for evaluating technology platforms and AI solutions, the adoption of such tools for visualizing data quality metrics, like Data Integrity and Data Coverage, remains unexplored.

B. *Total Cost of Ownership*

Total Cost of Ownership (TCO) in IT encompasses several cost components like (1) system design and infrastructure costs (the initial setup of data processing systems, computational resources, and storage), (2) maintenance and human resource costs (regular system upkeep, troubleshooting, and personnel expenses), (3) user operation costs (e.g., for database engines and business intelligence tools, this includes query execution time, latency, and handling approximate results [13]), and (4) costs of re-engineering, including costly redesigns of poorly modeled data systems when the user demands evolve.

In AI, ensuring high-quality data is a critical factor in the TCO of deploying models in business environments (see [14] for a robust classification of data quality costs). Poor data quality can significantly increase operational and business costs throughout the AI lifecycle. These costs manifest in several ways: (1) Low-quality data can lead to poorly trained models, requiring additional iterations of training and validation. This increases computational costs and prolongs deployment timelines. (2) Post-deployment, models operating on low-quality inference data are more likely to trigger monitoring alerts. These alerts necessitate frequent investigations, potentially leading to model re-tuning. (3) Errors stemming from data quality issues – whether in training data, inference data, or both – can result in business losses. Incorrect model outputs may harm customer satisfaction, operational efficiency, or decision-making accuracy, directly affecting the bottom line.

With large language models (LLMs) becoming a "hot topic," understanding their TCO is increasingly important. Measuring data quality for LLMs, including the evaluation of training and inference data, is an emerging challenge. The costs of maintaining high-quality data for such models are substantial, given their reliance on vast and diverse datasets.

Our previous research highlights the importance of diagnostic tools for AI models, as discussed in [15]. These tools help identify model errors, some of which may be rooted in data quality issues. Such diagnostics are valuable for pinpointing problems in both training and inference data. However, even the most advanced diagnostic systems have limitations and cannot identify all potential errors. Thus, investing in robust data quality analysis from the outset remains essential.

While poor data quality means significant problems, efforts to improve it are not without their own financial and operational implications. Within AI-infused data processing pipelines, additional costs of data enhancements emerge:

- External data. High-precision external data can be expensive, particularly for use cases requiring customer data, detailed measurements, or enriched metadata.
- Advanced parsing and quality enhancement tools. These tools improve data accuracy but at the same time, increase computational costs and latency.

- Human-involved data labeling and curation. Active learning and interactive tagging approaches, such as those explored in [16], involve human experts in data improvement processes. While effective, these methods vary in cost depending on the level of investment, such as using multiple experts for higher accuracy.

Effectively managing these costs is essential to optimizing the TCO for AI deployments, as both underinvestment and overinvestment in data quality can compromise the overall value and efficiency of AI solutions in applications.

The success of AI projects can be multifaceted, encompassing technical, ethical, and societal dimensions. Unlike traditional IT projects, AI initiatives involve unique challenges due to the complexity of algorithmic decision-making and its far-reaching impacts (see a recent study [17] for a review of AI success factors within the project management literature). A crucial metric for assessing the success of AI deployments is the return on investment (ROI), directly tied to the balance of investment in data quality and the value derived from AI solutions. Achieving a positive ROI depends on ensuring that the costs associated with improving Data Integrity and Data Coverage are justified by the benefits these improvements bring to AI performance and business outcomes.

Measuring ROI for AI involves evaluating quantifiable gains, such as cost reductions and revenue increases, and intangible benefits, including improved customer experiences, enhanced decision-making speed, and competitive positioning. Companies should monitor and adjust their data quality investments to ensure that the total cost of ownership is optimized, and the expected ROI is achieved or exceeded.

By visualizing data quality through QEDrants, business users will make informed decisions about which data sources to improve, ignore, or prioritize. This targeted approach helps organizations allocate resources efficiently, ultimately optimizing TCO. Once these decisions are made, systems (like e.g. BlueQuail developed by QED Software[3]) can operationalize them, offering guidance on feasible data improvement strategies. Additionally, integrating active learning techniques ensures a balance between data quality and human resource costs, optimizing the overall investment in data curation.

C. Data Governance and Security

Data governance is a critical yet expansive topic, often considered a cornerstone of effective data management. It encompasses the frameworks, policies, and procedures that ensure the data is managed as a valuable asset, aligning with organizational goals and regulatory requirements. Data governance is closely tied to data quality, as poor governance can lead to inconsistencies, inaccuracies, and compliance risks.

A key aspect of data governance is enabling business users to play an active role in data management. Traditionally, governance has been the domain of IT and data management professionals, but involving business stakeholders can bridge the gap between technical data policies and business needs. By equipping business users with tools like QEDrants, organizations can democratize data governance, allowing non-technical stakeholders to assess and influence data quality proactively.

Data security, though sometimes overlooked, is an equally important consideration in the context of AI and data quality. In business applications, security concerns frequently arise when sensitive data must be sent to external AI modules or third-party services. To mitigate the risks, organizations often anonymize or obfuscate the data before sharing it. However, this process can degrade data quality, introducing a trade-off between maintaining privacy and ensuring data reliability.

This trade-off was explored in [18], where the data was deliberately “corrupted” for business reasons, demonstrating the impact of security measures on data utility. Similarly, anonymization becomes particularly relevant when AI model development is outsourced to external firms, a common practice observed e.g. at QED Software[4]. Outsourcing can shift to crowdsourcing in competitive scenarios like those hosted on platforms such as knowledgepit.ai. A notable example is presented in [19], where sensitive communication data was stripped of its content to ensure privacy, rendering sentiment analysis infeasible. This highlights the broader challenge faced by all crowdsourcing platforms, including Kaggle, where data anonymization can limit the scope of achievable insights.

In both outsourced and crowdsourced AI projects, maintaining a balance between data security and data quality (which implies AI readiness) is crucial. Future iterations of the QEDrant framework may explore this trade-off, providing business users with visual tools to assess the impact of security measures on Data Integrity and Data Coverage.

IV. QEDrant Framework

This section lays the groundwork for understanding the QEDrant framework, focusing on its structure, core components, and core functionalities. We begin by addressing the foundational mechanics and metrics that drive the framework. Next, we shift to the user perspective, exploring how to interact with the framework and interpret its outputs. Finally, we revisit the foundations to consolidate key insights.

The QEDrant framework is a structured approach to assessing data quality, designed to help business users quickly understand the readiness of their data for AI applications. The framework organizes data assets into four quadrants based on two key metrics – Data Integrity and Data Coverage – allowing users to evaluate data reliability and completeness at a glance. The data quality analysis has a subsequent goal of recommending actions for data improvement or enrichment to better support AI. This is done without referring to any specific AI model. Instead, the framework provides foundational insights into data quality, helping users recognize the value and limitations of their data for future AI applications.

A. Data Integrity and Data Coverage

The QEDrant framework is grounded in established theories of data quality management, drawing on key metrics such as

[3] https://bluequail.ai/

[4] https://qedsoftware.com/

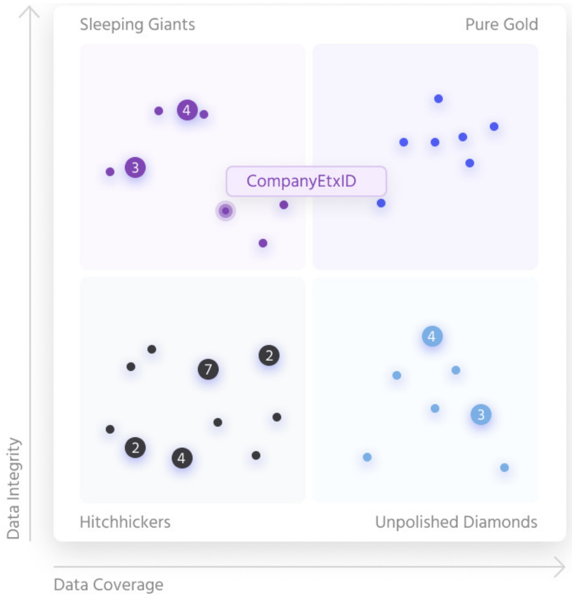

Fig. 2: QEDrant inspired by Gartner's Magic Quadrants^{TM}. It visualizes data assets based on Data Integrity and Data Coverage. Each quadrant (Sleeping Giants, Pure Gold, Unpolished Diamonds, Hitchhikers) categorizes the data due to its readiness and suitability for AI, providing an actionable overview for business users to prioritize data quality improvements.

accuracy, completeness, consistency, and relevance [4], [5]. As we already know, these dimensions are translated into two aggregate metrics within this framework: Data Integrity and Data Coverage. In the current version of the framework, these metrics are implemented for tabular data formats.

Data Integrity. Derived from the concepts of accuracy and consistency, it evaluates how closely the data aligns with expected semantic types and domain rules. It measures the reliability and error-free nature of the data, ignoring null values. More about the evaluation components:

- Data validity: Whether the data in a column adheres to defined business rules or domain constraints (e.g., no negative values in an age column).
- Data consistency: Consistent data format across the column (e.g., uniform date format). This metric is calculated per field and then aggregated across fields to provide a Data Integrity score for each table.

Data Coverage. It assesses whether the data is not too sparse. In the simplest version, one can think about it as the percentage of non-null values. However, a more advanced analysis of semantic types of missing values is required in future [20].

In Subsection IV-E, we will elaborate on how to estimate Data Integrity and Data Coverage in a more sophisticated way. However, we want to keep information as simple as possible for business users. Therefore, more advanced methods will need to come up together with their intuitive explanations.

B. Two Levels of Granularity

QEDrants operate across the following levels. Level 1 assesses the overall quality of data tables by aggregating metrics for Data Integrity and Data Coverage across columns. Level 2 goes deeper, analyzing individual columns in a table. These scores are visually represented in a QEDrant diagram (see Fig. 2), allowing users to quickly see where their data stands from the perspective of usefulness and reliability. This is a straightforward categorization of data quality.

Level 1: Table-Level Assessment. Each table is assigned a Data Integrity and Data Coverage score, derived by aggregating the basic metrics across all columns. This provides a high-level view of the table's suitability for AI. It enables users to prioritize tables for refinement or immediate application.

Level 2: Column-Level Assessment. For every column, the framework calculates its reliability and consistency based on validation criteria (e.g., adherence to semantic types or domain constraints). Intermediate outputs such as unique identifiers (UIDs), semantic types, and time-related columns are identified to provide a more granular understanding of data quality.

We refer to [21] for a vision of a richer hierarchy of granularities that can be useful for analyzing data quality.

C. QEDrant Representations

This subsection serves as a glossary for business users, providing a comprehensive overview of the QEDrant framework's key elements and functionalities. We explain the user interface (UI) features that make QEDrants accessible and actionable for non-technical stakeholders. Fig. 2 provides a visual reference. (For further study on visual navigation through QEDrants, we refer to [21] again.) We identify the following quadrants:

Pure Gold quadrant represents data with high integrity (accurate, reliable, consistent) and high coverage (comprehensive, minimum number of missing values). **Importance.** Pure Gold data is well-suited for high-stakes AI applications where accuracy and coverage are essential, such as predictive modeling and decision support systems. **Examples of Use.** Pure Gold data can be deployed immediately in AI, providing reliable insights with minimal risk of errors or biases.

Sleeping Giants represent high integrity but low coverage data, indicating that the data is accurate and consistent but has gaps or missing entries. **Importance.** This data may lack sufficient coverage for comprehensive analyses but is valuable in applications requiring precision and reliability within a limited scope. **Examples of Use.** Sleeping Giants are ideal for pilot AI projects or initial proof-of-concept models where accuracy is paramount, but complete coverage is not a requirement.

Unpolished Diamonds represent high coverage but low integrity data, suggesting the data is complete but may contain errors or inconsistencies. **Importance.** While suitable for exploratory analysis or feature discovery, Unpolished Diamonds require refinement before being applied to critical AI tasks. **Examples of Use.** This data can support early-stage analysis where the breadth of the data is valued over precision.

Hitchhikers represent low integrity and coverage, indicating the data that is both incomplete and potentially inaccurate or inconsistent. **Importance.** Hitchhikers are generally unsuitable for direct AI applications but could provide some value in non-critical exploratory tasks or after significant data enrichment and cleaning efforts. **Examples of Use.** This data may be useful for supplementary analyses, or in cases where additional cleaning can bring it up to a higher standard of usability.

D. Business Relevance

The QEDrant framework helps business leaders visualize data quality at a glance, empowering them to make informed decisions about AI readiness and data improvement tasks. By categorizing data assets into actionable quadrants, the framework enables organizations to:

- Leverage AI-ready assets (Pure Gold) immediately, maximizing ROI as discussed in Subsection III-B.
- Identify targeted analyses and data collection needs (Sleeping Giants) for precise insights.
- Prioritize data cleaning tasks (Unpolished Diamonds) to unlock valuable data potential.
- Avoid unnecessary costs on low-value data (Hitchhikers), optimizing TCO and resource allocation.

This approach not only clarifies data priorities but also aligns data quality efforts with business goals, reducing the barriers to effective AI adoption. For non-technical users, QEDrants provide a structured framework to guide data management strategies, helping them realize measurable business outcomes without deep technical expertise. This way, companies can confidently advance their data assets from potential to performance, setting a solid foundation for AI success.

E. More about Metric Derivations

Finally, let us go back to the problem of computing the Data Coverage and Data Integrity measures at particular levels of granularity. While the current version of QEDrants employs relatively simple methods, we acknowledge the potential for refinement and expansion. This subsection outlines a roadmap for improving these calculations, ensuring they better serve the practical goals described in subsequent sections.

Expanding to multimodal data. Future QEDrant releases should account for images, text, logs, etc. For such datasets traditional notion of a column does not apply. Instead, each modality (e.g., a camera feed, text document, sensor reading) may be treated as an independent "field." At Level 1, Data Coverage and Data Integrity may be computed separately for each modality, while Level 2 would aggregate them across modalities to provide a comprehensive view.

One possible approach is to transform raw multimodal data into intermediate vector or tensor representations. These representations, derived using appropriate tools for data transformation [22], may be evaluated using classical metrics. Multiple versions of these transformations might be sampled, with metrics averaged across them. These intermediate steps would remain hidden from users, unless they specifically request an explanation through an Explainable AI module.

Leveraging advanced learning techniques. To enhance metric accuracy, we propose more sophisticated, learning-based methods for estimating Data Coverage and Data Integrity:

- Feature selection approaches. Inspired by feature selection, we may dynamically generate hypothetical target variables based on the dataset's semantic context. For each target variable, quality measures for fields (or modalities) could be evaluated using filter-based or model-based methods. By averaging these results across various target variables, a more nuanced estimate of field quality may be obtained.
- Data and entity matching. Another promising direction involves leveraging a repository of historical datasets with validated Data Coverage and Data Integrity scores. By matching new datasets to similar historical datasets (using entity matching techniques), we can infer quality levels for new data. This approach would allow the system to "learn" from past data and apply those insights to new, incoming datasets.
- Interactive learning with expert feedback. Interactive learning can refine the framework. Starting with basic calculations, the framework can present edge cases to domain experts. Their feedback can be used to fine-tune the quality assessment models, gradually incorporating richer, more accurate metrics. Over time, these interactions can enable our framework to adapt and improve its recommendations. For a deeper discussion on active learning methodologies, see [16].

Towards continuous improvement. Even in production, user feedback plays a crucial role in improving the system. If business users disagree with QEDrant classification, their corrections can be fed back into the model for retraining, enabling continuous improvement. This feedback-driven approach ensures that Data Coverage and Data Integrity metrics evolve alongside the changing needs and contexts of the organization. By refining these metrics and incorporating advanced methods, QEDrants can provide more precise and actionable insights across a wide range of data types and use cases.

The above roadmap not only enhances the technical robustness of the framework but also ensures it remains adaptable to emerging challenges, such as multimodal datasets and evolving business requirements. While this framework is designed to analyze data independently of any specific AI application, future work can also extend its capabilities to recommend tailored data enrichment or cleaning actions based on AI project needs. Moreover, more advanced evaluation metrics and aggregation methods are envisioned in subsequent phases, aligning the framework more closely with specific AI objectives.

V. QEDrant Applications

By categorizing the data into four quadrants based on Data Integrity and Data Coverage, the QEDrant framework provides actionable guidance for each data asset. This section details practical use cases for each quadrant, illustrating how they can inform data strategies and improve decision-making. To give a comprehensive view, the section is structured as follows:

- Subsections V-A-V-D: Real-world examples of how each quadrant guides immediate business actions.
- Subsection V-E: Advanced scenarios – exploring how companies may push the boundaries of AI adoption.
- Subsection V-F: Integrations – embedding QEDrants into larger AI ecosystems to maximize their impact.

Each subsection is designed to help organizations leverage the QEDrant framework not only as a diagnostic tool but also as a driver for strategic improvements in data readiness.

A. Pure Gold

Practical Application: Data assets categorized as Pure Gold are immediately suitable for AI applications. This data can be confidently utilized in high-stakes AI initiatives such as predictive analytics, fraud detection, and customer segmentation.

Guidance for Use: The primary strategy with Pure Gold is to harness its rich and high-value data for strategic decision-making, focusing on areas where immediate, actionable insights can drive significant impact.

AI Readiness: Pure Gold data assets require minimal preparation. Their high quality supports reliable AI training, reducing the risk of errors and allowing for fast deployment.

Measurable Business Outcomes: Using Pure Gold data improves decision-making and operational efficiency. Examples:

- Customer segmentation: Accurate targeting enhances marketing ROI and customer engagement.
- Fraud detection: High data quality reduces false positives, minimizing financial losses.
- Predictive maintenance: Reliable performance data enables more accurate predictions, reducing downtime and optimizing resources.

For non-technical users, Pure Gold means “AI-ready” assets, allowing them to proceed with confidence and minimal effort.

B. Sleeping Giants

Practical Application: Sleeping Giants are accurate and reliable but incompleteness may restrict their usage in comprehensive analyses. They may be well-suited for limited or targeted analyses, where precision is more important than breadth.

Guidance for Use: The primary strategy is to leverage high integrity for targeted insights, but also identify areas where additional data collection could expand usefulness. For example, a retail company might use Sleeping Giants data to analyze customer behavior in a specific region, and then supplement it with new data collection efforts for broader insights.

AI Readiness: Sleeping Giants data is suitable for proof-of-concept models or analyses focused on precise questions within a limited scope. Organizations can proceed with smaller-scale AI initiatives, assessing the initial utility of the data while planning for future data enrichment.

Measurable Business Outcomes:

- Market analysis in targeted segments: Precise insights for specific demographics or regions, reducing the cost of large-scale data collection.
- Feature testing for product development: Using accurate but limited data to aid efficient R&D.

Non-technical users can leverage Sleeping Giants data for “targeted insights now, broader potential later.” This approach offers immediate value and provides a clear path for further data collection if more comprehensive analyses are desired.

C. Unpolished Diamonds

Practical Application: Unpolished Diamonds datasets can be regarded as comprehensive but in the same time they may contain errors or inconsistencies. This quadrant is ideal for identifying data cleaning tasks that can elevate its quality, making it suitable for more robust AI applications in future.

Guidance for Use: For Unpolished Diamonds, the focus should be on data cleaning and validation to improve data integrity. This includes tasks such as correcting inconsistencies, filling in missing values where possible, and standardizing data formats. Once cleaned, this data can transition to the Pure Gold quadrant, making it highly valuable for AI applications.

AI Readiness: Unpolished Diamonds are not immediately AI-ready but offer potential once data cleaning tasks are performed. These data assets can support exploratory analyses and feature discovery during the initial stages, but they should undergo refinement before being used in critical AI models.

Measurable Business Outcomes:

- Improved exploratory analysis: Cleaning the data enhances the reliability of trend analysis and feature discovery, making initial insights more trustworthy.
- Preparedness for advanced AI applications: Data cleaning converts Unpolished Diamonds into AI-ready assets, increasing the ROI of the data asset over time.

Non-technical users see Unpolished Diamonds as “the data with potential.” Data cleaning can unlock this potential, transforming these assets into reliable resources for AI.

D. Hitchhikers

Practical Application: Hitchhikers are unsuitable for immediate AI use. This data typically requires significant effort to clean and augment, which may not be worth the investment relative to its potential value.

Guidance for Use: Given their low quality, Hitchhikers should generally be deprioritized to avoid unnecessary costs. Limiting efforts on these data assets helps reduce the TCO associated with data preparation and maintenance. In cases where Hitchhikers data holds specific or supplementary value, it can be revisited for enhancement later, but for most business needs, focusing on other quadrants yields better returns.

AI Readiness: Hitchhikers data is generally not AI-ready and should not be prioritized for immediate usage. These assets can be kept as optional but are unlikely to directly support critical AI applications without substantial improvement.

Measurable Business Outcomes:

- Cost savings: By limiting efforts on Hitchhikers data, one can focus resources on higher-value data assets, reducing the TCO associated with data preparation.
- Focused data strategy: Deprioritizing low-quality data allows organizations to concentrate on assets that are

more likely to yield actionable insights, increasing the efficiency of data-related investments.

Hitchhikers are “not worth the investment right now.” Limiting resources spent on Hitchhikers helps streamline data strategy and focuses attention on more promising assets.

E. Making QEDrant-based Decisions

QEDrants provide users with data quality visualization, enabling them to make informed decisions impacting their data strategy and AI projects. While earlier subsections laid the groundwork for interpreting QEDrant diagrams, this part delves into more advanced scenarios where business stakeholders leverage QEDrant insights to drive key decisions. Below, we explore various decision-making contexts.

Investing in extra data sources. One common scenario is to identify gaps in Data Coverage or Data Integrity that could hinder the success of an AI project. For example, a dataset in the Sleeping Giants quadrant may signal the need for additional sources to improve coverage. Users may decide to:

- Purchase external datasets (e.g., market trends or demographic data).
- Enhance data collection (e.g., gathering more detailed customer feedback or expanding survey outreach).

Such investments aim at improving data completeness, ensuring that models trained on these datasets achieve broader applicability and higher performance.

Improving data generation and processing. Data assets classified as Unpolished Diamonds (high coverage, low integrity) suggest that while sufficient data exists, its quality is compromised by errors or inconsistencies. In this case, QEDrant analysis might prompt business users to:

- Enhance data parsing or transformation pipelines to improve accuracy.
- Implement stricter validation rules or automate error detection mechanisms.

For instance, a company relying on web-scraped data might identify parsing errors causing misclassification or duplication. Addressing these issues would improve data reliability, which in turn supports more robust AI models.

Reevaluating AI project goals. QEDrant insights can lead to strategic shifts in AI objectives. For example, if a dataset supporting an AI classification task is predominantly in the Hitchhikers quadrant, then business users may decide to:

- Reduce the task’s complexity, e.g. moving from 500 decision classes to 50 more generalized ones.
- Adjust accuracy expectations based on current data limitations, moving from a target of 95% accuracy to a more achievable 90%.

These adjustments allow for more realistic project goals, aligning expectations with the available data capabilities.

Deciding on project continuation or pivot. When a significant portion of critical datasets fall into problematic quadrants, such as Hitchhikers or Unpolished Diamonds, business users might face a more fundamental question: Is the project viable? Based on QEDrant insights, they may:

- Decide to halt the project until data quality improves.
- Pivot the project focus to areas where higher-quality data is available.

For example, an AI project initially designed to predict customer churn may be shifted toward identifying high-value customers if the churn-related data proves insufficient in quality.

Trade-offs and cost-benefit analysis. QEDrant diagrams also help users navigate trade-offs between data quality dimensions and project requirements. Consider the following:

- Time versus Quality: Should the project proceed with current data quality to meet deadlines, or is it worth delaying for data improvement efforts?
- Cost versus Accuracy: Would investing in high-quality data sources justify the incremental improvement in model performance?

These trade-offs are particularly relevant for projects where small quality gains come at a high cost, enabling business users to evaluate the ROI of data enhancement efforts.

Adjusting model complexity or evaluation metrics. Another scenario is to adjust the complexity of the AI model or the metrics by which its performance is evaluated. Examples:

- For datasets with lower integrity, shifting from precision-oriented metrics (e.g., precision/recall) to more robust metrics like F1-score or Matthews correlation coefficient might be advisable.
- Simplifying model architectures to reduce sensitivity to noisy or incomplete data, which can still provide actionable insights with reduced computational costs.

Collaboration and resource allocation. QEDrant insights also aid in optimizing cross-team collaboration. For instance, datasets requiring significant improvement might warrant additional resources from IT, data engineering, or third-party vendors. By identifying and prioritizing data assets based on their quadrant classification, organizations can allocate resources more effectively, focusing on the most critical datasets first.

In summary, QEDrants can provide a foundation for strategic decision-making across a variety of business and technical contexts. From data acquisition and pipeline optimization to project goal revision and resource allocation, QEDrant analysis empowers users to make data-driven choices that balance data quality, project feasibility, and business impact. Going further, we will explore specific examples of these advanced applications, demonstrating how QEDrant insights translate into actionable strategies for optimizing AI projects.

F. Deployments and Integrations

To maximize their utility, QEDrants must operate as part of a broader data ecosystem, seamlessly connecting with other modules and systems to drive actionable insights. This subsection explores how QEDrants can integrate with existing data infrastructure, support decision implementation, and potentially evolve into a recommendation engine capable of suggesting data quality improvements.

Interfacing with other modules. QEDrants should not function in isolation. Instead, they must interface with various components of the data pipeline, including:

- Data ingestion and transformation pipelines. Once a QEDrant analysis identifies data quality issues, the system can trigger automated data cleansing or transformation processes in connected ETL pipelines.
- Monitoring and diagnostic tools. QEDrants can feed data quality insights into monitoring systems to flag potential issues affecting model performance.
- AI model training modules. Data quality metrics provided by QEDrants can inform model training, helping select the most reliable datasets or identifying areas where synthetic data augmentation may be beneficial.

By integrating with these modules, QEDrants enable a feedback loop where data quality improvements translate directly into enhanced model performance and business outcomes.

Supporting decision implementation. A key aspect of QEDrants is to translate analysis into action. When users decide to address specific data quality issues – e.g. enhancing Data Coverage or correcting parsing errors – the framework should support seamless implementation. This can involve:

- Task automation. Automatically initiating data quality improvement tasks, such as filling missing values using imputation techniques or applying stricter validation rules during data ingestion.
- Workflow integration. Creating tickets in project management tools (e.g., Jira, Asana) to involve relevant teams (e.g., data engineering) in quality improving.
- Collaboration with external vendors. If external data sources are required, QEDrants can generate detailed procurement requirements based on identified gaps in Data Coverage or Data Integrity.

Recommendation engine potential. To further enhance its utility, the QEDrant framework may evolve into a recommendation engine, autonomously suggesting data quality improvement actions. This requires several key capabilities:

- Data-driven recommendations. QEDrants can analyze historical data quality improvement efforts and their outcomes, learning which interventions are most effective for specific types of data quality issues.
- External data. By integrating with external repositories and APIs, QEDrants can identify new data sources that may fill coverage gaps or improve data reliability.
- Predictive analytics. We can predict the potential impact of suggested improvements on AI performance and business gains, helping users prioritize actions.

Thus, by interfacing with other modules, supporting decision implementation, and evolving into a recommendation engine, QEDrants can not only identify data quality issues but also drive actionable, automated improvements.

VI. Conclusions and Future Directions

The QEDrant framework presents a structured approach for businesses to assess and prioritize their data assets in preparation for AI adoption. By categorizing data into four actionable quadrants, this model enables organizations to understand their data readiness for AI without requiring deep technical expertise. The framework guides non-technical users in identifying AI-ready data, determining areas for targeted analysis, prioritizing data cleaning tasks, and managing costs by deemphasizing low-value data assets.

The framework's strength lies in its ability to simplify complex data quality assessments, helping business leaders make informed decisions about data curation and improvement. This quadrant-based approach ultimately empowers companies to approach AI adoption with clarity and confidence. It lowers the barrier to AI by translating data quality dimensions into actionable insights for non-technical stakeholders, aligning data efforts with business goals. As a foundational step, QEDrants establish a roadmap for data quality improvement that can evolve with a company's AI maturity, supporting more advanced data strategies and AI applications over time.

As we continue to refine the QEDrant framework, several avenues for future work emerge, ranging from enhancing core functionality to exploring entirely new concepts. Below, we outline key directions for future R&D, some of which have been briefly mentioned in earlier sections.

Expanding data modalities. One significant area of future work is to extend QEDrants to support multimodal data. Currently, the framework focuses on tabular data, but many real-world AI applications rely on a mix of data types, including images, text, and sensor data. In future, we intend to:

- Develop methods for assessing Data Integrity and Data Coverage across different modalities.
- Introduce modality-specific metrics (e.g., resolution for images, semantic coherence for text).
- Enable users to view data quality metrics for individual modalities or entire multimodal datasets.

This expansion will introduce QEDrants to a wider range of industries, e.g., healthcare, autonomous systems, media.

Advanced methods for calculating data quality metrics. While the current approach to computing Data Integrity and Data Coverage relies on straightforward aggregation methods, more sophisticated techniques can improve accuracy and applicability. Potential directions of improvement include:

- Predicting data quality metrics, especially when direct calculations are infeasible or incomplete.
- Using dynamic target variables and feature selection to better evaluate the relevance of specific fields.
- Incorporating historical data repositories to infer quality metrics for new datasets based on their similarity to previously validated data.

These advancements would enable more precise assessments, especially for complex or evolving datasets.

Recommendations for data improvement. As discussed earlier, transforming QEDrants into a recommendation engine can significantly enhance their utility. Examples of future work:

- Developing algorithms that suggest targeted actions, such as acquiring new data sources, automatic data cleaning, or modifying AI project objectives.

- Integrating external data sources and metadata repositories to provide context-aware recommendations.
- Evaluating the impact of recommended actions through predictive analytics, helping users prioritize improvements based on expected business outcomes.

Interactive and adaptive learning. QEDrants may incorporate interactive learning mechanisms to continuously refine their assessments and recommendations. This may include:

- Collecting feedback from users on the accuracy and usefulness of QEDrant outputs.
- Employing active learning techniques to engage domain experts in reviewing edge cases, gradually improving the system's understanding of data quality.
- Implementing a closed-loop feedback system, where users' inputs directly influence future iterations.

Such capabilities would ensure that QEDrants remain responsive to user needs and evolving data environments.

Addressing trade-offs in data quality. Future work may also explore more nuanced trade-offs between data security and data quality, as already highlighted. For example:

- Investigating the impact of data anonymization and obfuscation on Data Integrity and Data Coverage.
- Developing visual tools to help users balance privacy and quality, potentially introducing new QEDrant variants focused on these trade-offs.
- Exploring real scenarios where such trade-offs are critical, e.g., outsourced / crowdsourced AI projects.

Real-time and dynamic data quality assessment. Another promising direction is to enable real-time data quality assessment, similarly to [23]. This would involve:

- Integrating QEDrants directly into live data pipelines to provide continuous monitoring and feedback.
- Developing dynamic visualization capabilities to reflect changes in data quality metrics over time.
- Supporting adaptive decision-making by alerting users to emerging issues, with immediate recommendations.

Real-time assessments may be particularly valuable in industries like finance, where timely insights are critical.

In summary, the QEDrant framework provides a strong foundation for data quality assessment, but its full potential lies in integration with broader ecosystems. By pursuing the outlined future work, we aim to make QEDrants even more versatile, precise, and user-friendly. Some further investigations, particularly related to more types of granularity levels and associated visualizations, can be found in [21].

VII. Acknowledgements

This research was co-funded by the Polish National Centre for Research and Development under the project MAZOWSZE/0198/19.

References

[1] U. Jagare, *Operating AI: Bridging the Gap Between Technology and Business*, Wiley, 2022.

[2] M. Świechowski, *The History of Artificial Intelligence: From Leonardo da Vinci to Chat-GPT*, Amazon KDP, 2024.

[3] G.L. Geerts and D.E. O'Leary, "V-Matrix: A Wave Theory of Value Creation for Big Data," *International Journal of Accounting Information Systems*, vol. 47, pp. 100575, 2022.

[4] R.Y. Wang and D.M. Strong, "Beyond Accuracy: What Data Quality Means to Data Consumers," *Journal of Management Information Systems*, vol. 12, no. 4, pp. 5–34, 1996.

[5] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino, "Methodologies for Data Quality Assessment and Improvement," *ACM Computing Surveys*, vol. 41, pp. 16:1–16:52, 2009.

[6] S. Sadiq and M. Indulska, "Open Data: Quality Over Quantity," *International Journal of Information Management*, vol. 37, no. 3, pp. 150–154, 2017.

[7] Y. Gil and B. Selman, "A 20-year Community Roadmap for Artificial Intelligence Research in the US," *AI Magazine*, vol. 40, no. 1, pp. 8–24, 2019.

[8] R.Y. Wang and S.E. Madnick, "A Polygen Model for Heterogeneous Database Systems: The Source Tagging Perspective," in *Proceedings of VLDB 1990*, 1990, pp. 519–538.

[9] L. Pipino, Y.W. Lee, and R.Y. Wang, "Data Quality Assessment," *Communications of the ACM*, vol. 45, pp. 211–218, 2002.

[10] M.G. Kahn, T.J. Callahan, J. Barnard, A.E. Bauck, J. Brown, B.N. Davidson, H. Estiri, C. Goerg, E. Holve, S.G. Johnson, S.T. Liaw, M. Hamilton-Lopez, D. Meeker, T.C. Ong, P. Ryan, N. Shang, N.G. Weiskopf, C. Weng, M.N. Zozus, and L. Schilling, "A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data," *Journal of Electronic Health Data and Methods*, vol. 4, no. 1, pp. 18, 2016.

[11] Y. Lee, D. Strong, B. Kahn, and R. Wang, "AIMQ: A Methodology for Information Quality Assessment," *Information & Management*, vol. 40, pp. 133–146, 12 2002.

[12] S. Bresciani and M.J. Eppler, "Case Nr.2, 2008 – Updated in 2010 Gartner's Magic Quadrant and Hype Cycle," 2010.

[13] M. Kowalski, D. Ślęzak, and P. Synak, "Approximate Assistance for Correlated Subqueries," in *Proceedings of FedCSIS 2013*, 2013, pp. 1455–1462.

[14] M. Eppler and M. Helfert, "A Classification and Analysis of Data Quality Costs," in *Proceedings of ICIQ 2004*, 2004, pp. 311–325.

[15] A. Janusz, A. Zalewska, Ł. Wawrowski, P. Biczyk, J. Ludziejewski, M. Sikora, and D. Ślęzak, "BrightBox – A Rough Set Based Technology for Diagnosing Mistakes of Machine Learning Models," *Applied Soft Computing*, vol. 141, pp. 110285, 2023.

[16] D. Kałuża, A. Janusz, and D. Ślęzak, "Robust Assignment of Labels for Active Learning with Sparse and Noisy Annotations," in *Proceedings of ECAI 2023*. 2023, vol. 372 of *Frontiers in Artificial Intelligence and Applications*, pp. 1207–1214, IOS Press.

[17] G.J. Miller, "Artificial Intelligence Project Success Factors – Beyond the Ethical Principles," in *Post-Proceedings of FedCSIS-AIST 2021*. 2021, vol. 442 of *Lecture Notes in Business Information Processing*, pp. 65–96, Springer.

[18] M.S. Szczuka, A. Janusz, B. Cyganek, J. Grabek, Ł. Przebinda, A. Zalewska, A. Bukała, and D. Ślęzak, "IEEE BigData Cup 2022 Report Privacy-preserving Matching of Encrypted Images," in *Proceedings of IEEE BigData 2022*. 2022, pp. 6471–6480, IEEE.

[19] A. Janusz, G. Hao, D. Kałuża, T. Li, R. Wojciechowski, and D. Ślęzak, "Predicting Escalations in Customer Support: Analysis of Data Mining Challenge Results," in *Proceedings of IEEE BigData 2020*. 2020, pp. 5519–5526, IEEE.

[20] T. Mroczek, D. Gil, and B. Pękała, "Fuzzy and Rough Approach to the Problem of Missing Data in Fall Detection System," *Fuzzy Sets and Systems*, vol. 480, pp. 108868, 2024.

[21] A. Powała and D. Ślęzak, "Hierarchical Approach to Data Quality Understanding in QEDrant Framework," in *Proceedings of IEEE BigData 2024*. 2024, IEEE.

[22] M. Bartoszuk, J. Litwin, M. Wnuk, and D. Ślęzak, "Tensor-based Approach to Big Data Processing and Machine Learning," in *Proceedings of IEEE BigData 2022*. 2022, pp. 6188–6194, IEEE.

[23] J. Bicevskis, Z. Bicevska, A. Nikiforova, and I. Oditis, "Towards Data Quality Runtime Verification," in *Proceedings of FedCSIS 2019*, 2019, vol. 18 of *Annals of Computer Science and Information Systems*, pp. 639–643.