UNIVERSITY OF TARTU
Institute of Computer Science
Data Science Curriculum

Kaisa Käosaar

Exploring Data Quality Management Challenges and the Emerging Role of AI Solutions

Master's Thesis (15 ECTS)

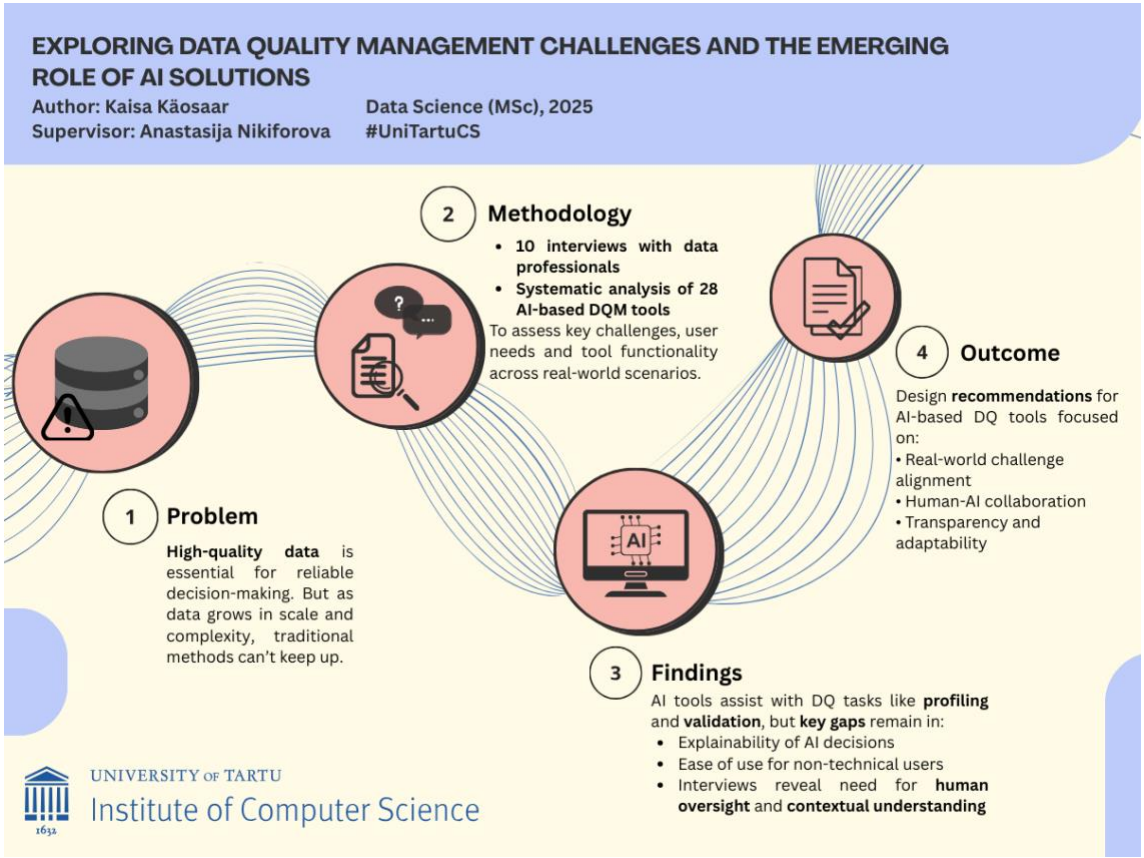
Supervisor:
Anastasija Nikiforova, PhD

Tartu 2025

Exploring Data Quality Management Challenges and the Emerging Role of AI Solutions

Abstract:

High-quality data is essential for reliable decision-making and efficient operations across organizations. However, managing data quality (DQ) remains a complex and resource-intensive challenge. In response, artificial intelligence (AI) has been increasingly integrated into data quality tools. Yet, there is limited understanding of whether these AI-powered tools meet the practical needs of data professionals. This study addresses this gap by investigating how current AI-enabled data quality tools address the practical needs and challenges faced by data professionals. Using a mixed-methods approach, the study combines semi-structured expert interviews with a structured analysis of 28 AI-enabled data quality tools. Interview findings reveal persistent challenges such as limited support for unstructured data, low explainability, fragmented workflows, and minimal involvement of business users. While many tools perform well in data profiling, rule-based validation, and structured data integration, fewer support collaboration, domain-specific customization, or transparent AI behaviour. Despite progress, most tools fall short of meeting the complex and context-driven demands of enterprise-level data quality management (DQM).



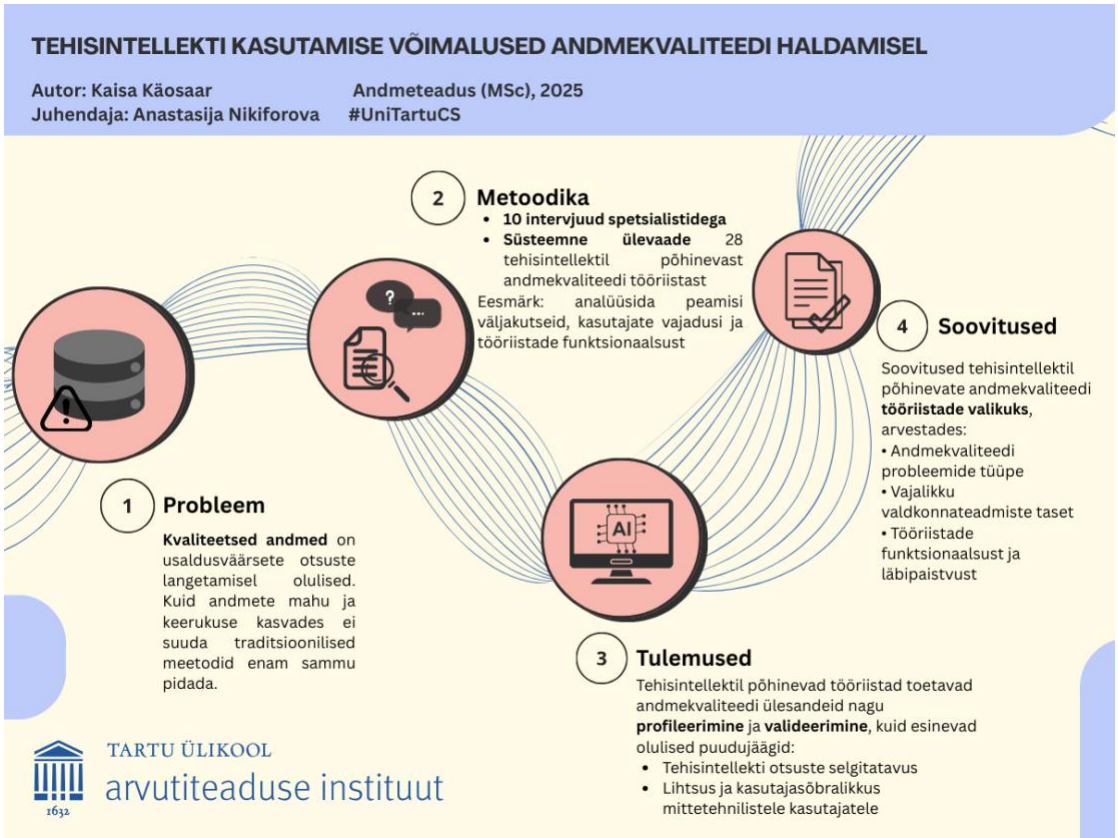
Keywords: Data quality, data quality management, artificial intelligence

CERCS: P175 Informatics, systems theory, P176 Artificial Intelligence

Tehisintellekti kasutamise võimalused andmekvaliteedi haldamisel

Lühikokkuvõte:

Kvaliteetsed andmed on eelduseks usaldusväärsete otsuste tegemisel ja tõhusa tegevuse tagamisel igas organisatsioonis. Andmekvaliteedi haldamine on aga jätkuvalt keeruline ja ressursimahukas väljakutse. Järjest enam on hakatud integreerima tehisintellekti (TI) andmekvaliteedi tööriistadesse. Siiski on piiratud arusaam sellest, kas need TI-põhised tööriistad suudavad vastata andmespetsialistide praktilistele vajadustele. Käesolev töö käsitleb, kuidas praegused TI-toega andmekvaliteedi tööriistad vastavad andmespetsialistide vajadustele ja väljakutsetele. Uuringus kasutatakse kombineeritud metoodikat - poolstruktureeritud ekspertintervjuud ja 28 AI-toega andmekvaliteedi tööriista struktureeritud analüüs. Intervjuude tulemused toovad esile mitmeid probleeme, sealhulgas piiratud tuge struktureerimata andmetele, madalat selgitatavust, killustatud töövooge ning vähest ärikasutajate kaasatust. Kuigi paljud tööriistad toimivad hästi andmete profileerimisel, reeglipõhisel valideerimisel ja struktureeritud andmete integreerimisel, toetavad vähesed neist koostööd, valdkonna spetsiifilist kohandamist või läbipaistvat TI-käitumist. Vaatamata edusammudele ei ole enamik tööriistu veel kooskõlas spetsialistide tegelike vajaduste ja andmekvaliteedi haldamise kontekstitundliku keerukusega.



Keywords: Andmekvaliteet, andmekvaliteedi haldamine, tehisintellekt
CERCS: P175 Informaatika, süsteemiteooria, P176 Tehisintellekt

Table of Contents

Introduction

In an increasingly data-driven world, organizations rely on high-quality data to make informed decisions, comply with regulations and gain competitive advantage (Wang & Strong, 1996; Kavak & Rusu, 2025). As data volume, variety and velocity continue to grow, ensuring that data remains dependable has become a complex and resource-intensive task (Batini et al., 2009). Traditional data quality management methods, often based on manual processes and rule-based systems, struggle to keep pace with the scale and diversity of modern data environments (Batini et al., 2009).

At the same time, advances in artificial intelligence have introduced new possibilities for automating and enhancing DQM practices. AI-empowered solutions can detect anomalies, suggest validation rules and support real-time monitoring at a scale that was previously unmanageable (Gami et al., 2024; Zhang et al., 2023). While the potential of AI in this domain is recognized, the extent to which these tools meet the actual needs and constraints of data professionals remains underexplored. Questions around explainability, usability and integration into existing workflows continue to pose practical barriers (Hardinges et al., 2024; Whang et al., 2023).

The objective of this thesis is to explore the challenges faced by data experts in managing data quality and to identify opportunities for AI to support these efforts. Through expert interviews, the study uncovers pain points in current DQM approaches and expectations for AI integration. Based on these insights, the study examines how well AI-empowered tools claim to address these needs, focusing on their stated functionalities and alignment with practitioner priorities.

The research questions (RQ) were developed based on an initial review of the literature on data quality management, as well as early discussions with data professionals. The goal was to frame the study in a way that captures both the practical challenges faced by data experts in DQM and the role of AI in addressing these issues. The following research questions were designed to guide the data collection, analysis process and interpretation of findings throughout the study:

RQ1: What are the most common challenges faced by data professionals in managing data quality?

The study aims to understand the pain points that data quality experts face in DQM, particularly those that AI-empowered tools are expected to help address.

RQ2: How do currently available AI-empowered data quality tools address the practical challenges faced by data professionals?

This question was designed to identify existing AI-empowered tools and assess the extent to which their features align with the real-world challenges raised by data quality practitioners. Through a systematic search and comparison with expert insights, the study seeks to understand whether these tools align with real-world DQM needs.

To achieve this, the study adopts a mixed-methods approach, combining qualitative insights from expert interviews with a systematic analysis of AI-empowered data quality tools. The interviews provide a practitioner's perspective on common challenges, while the tool analysis examines the extent to which current offerings reflect these expectations. By comparing expert-identified needs with the stated capabilities of current tools, the study explores how well existing AI-empowered solutions align with the practical challenges of data quality management.

The thesis is structured as follows: the next section presents an overview of the related works. This is followed by the theoretical background on data quality - its definition and dimensions, as well as the emerging role of AI and generative AI in this domain. The third section is the methodology chapter, which outlines the interview design, tool search strategy and evaluation protocol. The subsequent chapters present the findings from the interviews and tool assessment, followed by a discussion of their implications. The thesis concludes with a summary of the main contributions and recommendations for future research and practice.

This thesis made use of OpenAI's ChatGPT[1] for grammar refinement, wording and structural suggestions.

[1] OpenAI. (2025). ChatGPT (May 13 version) [Large language model]. https://chat.openai.com/

1. Theory and Background

This section provides the foundational concepts and literature relevant to the study of data quality management and the emerging role of AI in this field. It introduces key data quality dimensions, challenges in managing data quality and current AI techniques applied to enhance data quality management. This theoretical groundwork establishes the context for the subsequent sections.

1.1 Importance of Data Quality

Data has become one of the most valuable assets for organizations. However, data is only a valuable asset if it is reliable and trustworthy (Redman, 2001). Organizations across industries rely on data to gain insights, automate processes, and make informed decisions. However, the effectiveness of these activities is directly connected to the quality of the underlying data (Batini et al., 2009; Serra et al., 2024).

While organizations increasingly recognize the importance of data quality, in practice, many still rely on internal system data without properly managing its quality (Jiang & Zhao, 2012). Recent findings by Haruki et al. (2025) confirm that despite growing attention to data quality, the evaluation processes remain inconsistent and often depend heavily on informal judgments, underscoring that this challenge persists today (Haruki et al., 2025).

Although the concept of data quality seems straightforward, in practice, there isn't a universal standard, definition or measurement for it. Wang and Strong defined the term as “data that are fit for use by data consumers” (Wang & Strong, 1996). Scannapieco and Catarci (2002) have defined it through characteristics - “a set of characteristics that data should own, such as accuracy, i.e. a degree of correctness, or currency, i.e. a degree of updating”.

Data quality can be evaluated using data quality dimensions. The idea was first proposed by Wang and Strong (1996). They defined dimensions as a “set of data quality attributes that represent a single aspect or construct of data quality” (Wang & Strong, 1996, p. 6). In their study, they defined 15 dimensions to categorize these attributes. Figure 1 presents the conceptual framework of data quality proposed by Wang and Strong, which categorizes data quality dimensions into four key groups: intrinsic, contextual, representational, and accessibility.

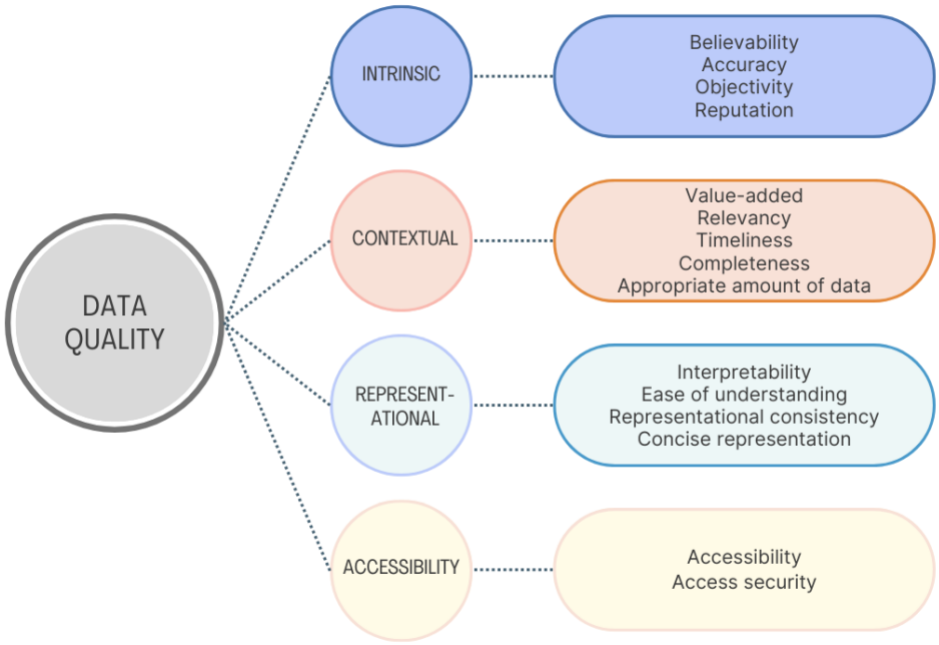


Figure 1. *Conceptual Framework of Data Quality.* Adapted from Wang and Strong (1996).

Since Wang and Strong's original work, other researchers have proposed alternative sets of dimensions and refined the definitions of existing ones. For example, Pipino, Lee, and Wang expanded on the original framework by identifying 16 dimensions, including attributes such as 'free-of-error' and 'concise representation' (Pipino et al., 2003). Redman emphasized the practical application of data quality dimensions in business contexts, highlighting the importance of accuracy, timeliness and completeness in operational decision-making (Redman, 2001). International standards such as ISO/IEC 25012:2008 similarly define a set of data quality characteristics, distinguishing between inherent and system-dependent data quality attributes (ISO/IEC 25012, n.d.).

While Wang and Strong's (1996) framework remains influential, data quality dimensions are not universally defined, and different domains and industries have adapted them based on their specific data environments and business requirements (Pipino et al., 2002; Serra et al., 2024). Most studies consistently identify accuracy, completeness, consistency and timeliness as core dimensions. According to Pipino et al. (2002), these dimensions are defined as follows:

- **accuracy** - the extent to which data correctly describes the real-world values it is intended to describe;
- **completeness** - the measure of whether all necessary data is present and provides sufficient detail for its intended use;

- **consistency** - the extent to which data remains uniform and does not conflict across different sources or systems;
- **timeliness** - the measure of how up-to-date the data is and whether it is available when needed.

Researchers have developed structured frameworks to conceptualize and manage data quality. These models aim to provide a standardized approach to assessing, improving and maintaining data quality across different types of data and business contexts. One of the most widely referenced frameworks is Total Data Quality Management (TDQM). The TDQM framework was introduced by Wang (1998) as a systematic approach to managing data quality in organizations. It builds on the principles of Total Quality Management (TQM), which emphasizes continuous improvement and process control. TDQM defines data quality management as a cycle consisting of four key phases:

- **definition** – identifying business requirements and designing the information product, including data quality attributes and models;
- **measurement** – defining metrics and assessing current data quality to uncover issues;
- **analysis** – investigating the root causes of identified data quality problems;
- **improvement** – selecting and applying strategies to address issues, feeding into the next cycle.

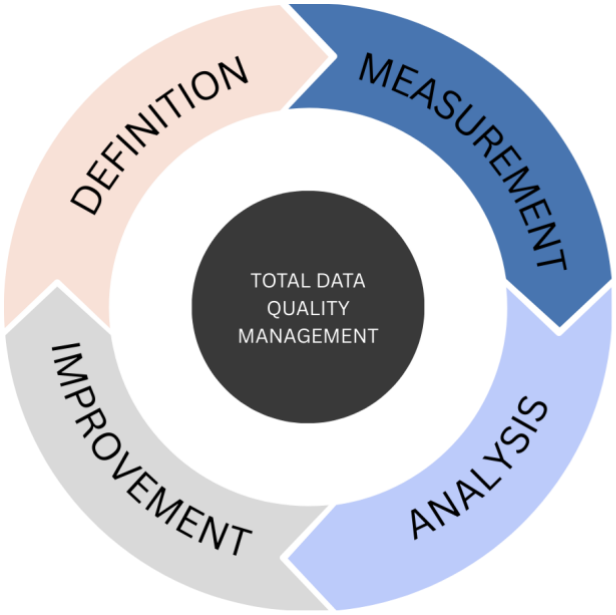


Figure 2. Total Data Quality Management Cycle. Adapted from Wang, R. Y. (1998).

The cyclical nature of TDQM highlights the importance of treating data quality management as an ongoing process rather than a one-time effort. This aligns with business practices where data is continuously collected, processed and analysed (Wang, 1998).

Despite these frameworks, ensuring high data quality remains a complex challenge due to the subjective nature of the concept. Whether data is qualified as high-quality depends largely on its intended use and the specific needs of users (Batini et al., 2009). A dataset considered

complete and accurate in one context may be inadequate in another, highlighting the importance of contextual evaluation in data quality assessment (Wang & Strong, 1996).

1.2 Consequences of Poor Data Quality

Data quality can be affected at any stage of data handling, including data generation at sources, integration and profiling activities, data transformation processes and database modelling. Errors may arise during extraction, transformation, and loading (ETL) procedures, as well as during storage and access phases (Ranjit & Kawaljeet, 2010).
The cost of poor data quality is most immediately felt through financial losses. Research by Gartner in 2020 indicates that organizations attribute an average annual loss of at least $12.9 million to poor data quality (*Data Quality*, n.d.). Poor data quality can also cause operational inefficiencies, leading to wasted resources and delays, as well as customer dissatisfaction, which can damage trust and brand reputation (Wang et al., 2024).
Moreover, it can expose organizations to regulatory penalties. A notable example is Citigroup, which was fined $136 million by U.S. bank regulators in 2024 for failing to address data management deficiencies identified in 2020 (Price et al., 2024). The bank was cited for making "insufficient progress" in improving its data governance and risk management practices, highlighting the regulatory risks associated with poor data quality.

Achieving complete consistency and error-free data, “absolute data quality”, is generally considered unattainable, especially in large, dynamic datasets (Nikiforova, 2020). As organizations process and integrate data from multiple sources, data quality issues inevitably arise. They can emerge due to factors such as source integration challenges, schema mismatches, data migration, and human error (Batini et al., 2009). While data can never be entirely error-free, a certain degree of data quality issues is often considered acceptable (Redman, 2001). Thus, organizations should aim for fit-for-purpose data quality, where data meets the specific needs of users while minimizing critical defects (Pipino et al., 2003).

1.3 Challenges in Data Quality Management

Managing data quality remains one of the most complex and resource-intensive tasks for organizations (Batini et al., 2009). As the volume, variety and velocity of data continues to grow, traditional approaches to data quality management face increasing limitations (Shah et al., 2024).

While automation and AI tools are increasingly used to detect and resolve data quality issues, they often fall short when it comes to understanding the context behind the data. Business domain knowledge remains essential for interpreting and managing data quality effectively (Holstein et al., 2024).

For example, AI might flag an outlier in a financial dataset as an error, when in reality it reflects a legitimate business transaction — such as a currency revaluation. As anomaly detection should support expert judgment, such discrepancies risk being misinterpreted without domain input (Vilella et al., 2024).

Data quality rules are formalized expressions of expectations that define acceptable data values based on business requirements and operational logic (Loshin, 2011). These rules may address aspects such as completeness, validity or consistency across systems. Defining such rules typically requires collaboration between technical teams and business stakeholders (Loshin, 2011). These rules are often not universal but highly specific to the organization, process or industry in question (Abraham et al., 2019).

Modern data environments consist of a variety of data types that differ in structure, format and complexity. Each data type presents unique challenges for data quality management. Structured data can often be validated using predefined rules and constraints, making it relatively easier to profile and cleanse. However, semi-structured and especially unstructured data pose greater challenges (De Haan et al., 2024). The lack of formal schema and inconsistencies in format can lead to difficulties in measuring data quality dimensions (Batini et al., 2009).

Moreover, organizations increasingly deal with real-time or streaming data, especially in sectors like finance or logistics. These high-velocity data streams often include mixed data types and demand immediate validation, leaving little room for manual inspection. The dynamic nature of such data makes it more prone to errors and complicates efforts to establish and maintain reliable data quality controls (Cichy & Rass, 2019).

While structured data allows for relatively straightforward quality checks, the growing prevalence of semi-structured, unstructured and real-time data significantly increases the complexity of data quality management. This shift necessitates more advanced, often AI-assisted methods for profiling, cleansing and validating data across diverse formats.

The way data is stored and organized has a significant impact on data quality management. Relational databases are typically governed by strict schemas and constraints, which facilitate data validation and consistency (Harrington, 2016). However, as organizations transition to more flexible and scalable storage systems, which allow the storage of vast amounts of raw, unstructured and semi-structured data—the risk of data quality issues increases.

In modern organizations, often data is no longer confined to traditional relational databases. Instead, it is distributed across a wide range of storage architectures, including data warehouses, data lakes and data mesh environments. Each of these architectures introduces distinct challenges for maintaining data quality (Siddiqa et al., 2017). Data lakes often lack the same level of schema enforcement, leading to problems such as duplicate records, missing metadata and inconsistencies across datasets (Nargesian et al., 2019). Moreover, decentralized models like data mesh complicate efforts to maintain a unified view of data quality. When data is managed by decentralized domain teams, there is often variation in how quality standards are defined and enforced. Without centralized governance or coordination mechanisms, it becomes difficult to track lineage, enforce validation rules or ensure consistency across domains.

Another challenge lies in the integration of data from different systems and formats. As data is ingested from various sources into centralized or federated storage environments,

inconsistencies in formatting, data definitions and update frequencies can lead to misalignment and reduced reliability of the data.

Ensuring high data quality is a continuous process that requires a combination of structured governance, validation techniques and automated tools. Organizations must implement effective strategies to identify quality issues early, maintain consistency across datasets and ensure that data remains reliable for decision-making. Traditionally, data quality has been managed through manual processes, predefined rule-based validation and oversight from data professionals. As the volume and complexity of data continue to grow, these methods are no longer sufficient on their own. Traditional approaches to data quality management typically rely on manual processes, which are both time-consuming and prone to human error. This shift has led to growing interest in the use of artificial intelligence to support data quality management.

1.4 The Role of AI and Generative AI in Data Quality Management

Artificial Intelligence refers to a system's ability to correctly interpret external data, learn from it and use those insights to achieve specific goals through flexible adaptation (Kaplan & Haenlein, 2019).

Due to the increasing scale and complexity of data, AI is emerging as a powerful solution to automate and enhance data quality management. AI-driven techniques can detect patterns, identify errors and even predict potential data quality issues before they impact business operations. Unlike rule-based systems, AI can learn from past corrections and adapt to new data structures (Gami et al., 2024). In the context of DQM, AI introduces several techniques that bring advantages over traditional methods:

- anomaly detection – AI models can identify inconsistencies and outliers in large datasets more effectively than rule-based systems (Gami et al., 2024);
- error correction – machine learning models can predict errors based on learned patterns (Gami et al., 2024);
- automation – AI can automate repetitive data validation tasks, freeing up human resources for higher-level activities (Zhang et al., 2023);
- predictive models – AI can anticipate future data quality issues by analysing historical patterns (Bauskar, 2024).

As artificial intelligence continues to evolve, one of the most significant recent developments is the emergence of Generative AI (GenAI). Building upon traditional AI techniques, GenAI enables systems to not only analyse and learn from data but also to generate new content, predictions or transformations based on learned patterns (Dhoni, 2023). This evolution opens new possibilities for applications in data quality management. Generative AI introduces new possibilities for improving data quality (Azeroual, 2024):

- data enrichment – filling missing values or suggest alternatives based on learned patterns;

- consistency checks – AI can identify and correct inconsistencies across different datasets;
- context-aware validation – generative AI models can assess data quality within the context of its intended use, enhancing accuracy and relevance.

While AI offers significant potential for improving data quality management, it also presents certain limitations.

Explainability in artificial intelligence refers to the extent to which the internal processes and decision-making pathways of a model can be understood and interpreted by humans (Angelov et al., 2021). While training data transparency contributes to the explainability of AI systems, true explainability also depends on model architecture, interpretability techniques and algorithmic transparency. Without insight into what data models are trained on, it becomes difficult to evaluate the fairness, safety or legal compliance of these systems (Shahzad et al., 2025).

These systems often function as black boxes, making it difficult to trace outputs back to specific data points or training inputs (Zednik, 2021). While explainability challenges can sometimes be mitigated in traditional AI systems through techniques such as model visualization or feature attribution, overcoming opacity in generative AI models is more difficult (Hardinges et al., 2024). This is due to the vast complexity of generative architectures, the probabilistic nature of output generation and the often unknown or untraceable origins of training data, which collectively make it challenging to interpret or validate how specific outputs are produced (Zhang & Zhang, 2025). This lack of transparency can undermine efforts to audit data lineage, identify sources of bias, or validate the quality of AI-generated content and the reliability of the insights it produces (Zednik, 2021; Hardinges et al., 2024).

Additionally, GenAI models are prone to hallucinations - plausible-sounding but inaccurate or misleading outputs, which can contaminate downstream data processes and result in flawed business insights or decision-making (Shahzad et al., 2025).

Another critical limitation involves data privacy and confidentiality. When employees interact with AI tools hosted by third parties, sensitive information may be exposed, raising security and compliance risks (Véliz et al., 2024). These risks are particularly acute in sensitive domains such as healthcare. Bak et al. (2022) note that privacy concerns complicate the balance between protecting patient data and enabling effective AI-driven data quality management. Furthermore, limited transparency around data use and governance in healthcare AI creates significant challenges for maintaining data quality while respecting regulatory requirements.

From a regulatory standpoint, the use of generative AI also raises concerns around liability and compliance. As AI begins to influence professional outputs, organizations may find themselves responsible for incorrect or unsafe outcomes produced by these systems (Buiten, 2019).

Beyond regulatory concerns, the adoption of AI-empowered data quality solutions also faces practical challenges. Whang et al. (2023) highlight that technical complexity can pose barriers for organizations with limited AI expertise. Furthermore, integrating AI tools into existing data infrastructures and workflows remains a significant hurdle, particularly when legacy systems or fragmented architectures are involved. Addressing these operational issues is crucial to ensuring that the potential benefits of AI for data quality management can be realized effectively (Whang et al., 2023).

These limitations underscore that while AI and GenAI offer promising enhancements to data quality management, their adoption must be approached with caution. Responsible integration requires not only technical innovation but also robust oversight, strong organizational trust and alignment with real-world constraints (Daly et al., 2025).

1.5 Related Works

This section gives an overview of prior studies that analyse data quality tools and practices, which offer relevant context for understanding the broader landscape of DQM tool capabilities. This section gives overview of three works that provide context for the present study.

Zhou et al. (2024) conducted a survey on data quality dimensions and available tools for machine learning contexts. They reviewed core data quality dimensions, metrics and key functions of 17 tools, highlighting their strengths and limitations for data-centric AI. While their work thoroughly addresses challenges in applying data quality tools to ML, it remains focused on ML-specific contexts and does not include empirical evaluation or practitioner perspectives on AI tool effectiveness.

Ehrlinger and Wöß (2022) conducted a systematic survey identifying 667 data quality software tools, from which 13 (8 commercial and 5 open-source) were selected for detailed analysis based on functionality in data profiling, quality measurement, and automated monitoring. Their analysis focuses primarily on data quality tool functionalities and does not specifically limit its scope to AI-empowered tools or address recent advances in AI-enabled features.

Tamm et al. (2025) conducted a study focused on automated rule detection for data quality management in data warehouse environments. The research included a systematic review of 151 data quality tools from both academic and industry sources, aiming to assess their capabilities for rule detection and enforcement. The findings revealed that most tools focused on data cleansing within domain-specific systems, with only a small subset supporting automated rule discovery. They primarily addressed rule inference, aspects of AI were not a part of this analysis.

These studies contribute important knowledge regarding tool functionalities, however none of them explicitly address the intersection of AI-empowered tools with practitioner perspectives. The objectives of these prior studies differ from the present research, which combines expert interviews with a structured analysis of AI-empowered data quality tools.

2. Methodology

This thesis uses a mixed-methods approach. The study combines a systematic analysis of AI-powered data quality tools with semi-structured expert interviews to provide an understanding of both the capabilities of existing tools and the perspectives of data professionals. The tool identification and screening process was guided by principles adapted from the systematic mapping review methodology proposed by Kitchenham et al. (2013). This section outlines the research design and methods used in the study, including the search strategy for identifying relevant tools, the criteria used for selection and the framework for evaluating them. It also describes the process for designing and conducting the interviews and the approach for analysing the collected data.

2.1 Interviews with Experts

This section gives an overview of the qualitative study methodology conducted through expert interviews. It outlines the interview design, protocol, participant selection, and analysis methods used to gather insights on data quality management and AI adoption.

2.1.1 Interview Design

This study used semi-structured in-depth interviews, consisting of open-ended questions. This format was chosen because it provides a structured framework for comparing insights across participants while also allowing flexibility to explore emerging topics during the conversation (DiCicco-Bloom & Crabtree, 2006). This approach ensured that key themes were consistently addressed while enabling participants to share context-specific insights and uncover new perspectives that may not have been anticipated during the question development phase (Rutledge, 2020). The findings from the interviews directly informed the development of the tool evaluation protocol. Interviews were conducted in English via Microsoft Teams and lasted approximately 60 minutes. Audio recordings were made with participant consent and were later transcribed.

2.1.2 Interview Protocol

The interview protocol was designed to explore the challenges faced by data professionals in managing data quality and to examine the role and limitations of AI in data quality management. The protocol was developed based on the research questions and themes covered in the literature review, particularly the gaps in AI-driven DQM practices and the complexity of current DQ challenges (see Section 2). The protocol was shared with participants ahead of the interview and is available in Appendix I. Participants were also provided with a written informed consent form that included information about data usage, anonymity and withdrawal rights. The protocol consisted of five main sections, beginning with general background questions, followed by more detailed questions about current practices, the use of AI and future expectations.

The first section intended to examine the participant's professional background, experience level and the data domain in which they work with data quality management (e.g. finance, healthcare). The second section explored the challenges data professionals face in managing data quality. It addressed issues related to data types, data structures, quality dimensions and resource constraints. The third section focused on how data quality is currently managed, including the degree of automation, types of tools used and how business knowledge is integrated into the process. The next section aimed to explore the role of AI in data quality management, examining whether participants currently use AI tools, in which areas they find AI most helpful and what issues might prevent wider adoption. The final section focused on expectations for the future of DQM, particularly regarding generative AI and whether participants believed it could shift how organizations manage and govern their data. All interview questions are listed in Table 1 below.

Table 1. Interview Sections and Questions.

Interview Protocol Section	**Interview Questions**
General	- What is your area of expertise? - How long have you been working with data quality management? - What are the DQM tasks you are typically involved in and for whom (e.g., businesses, governments, internal teams)? - Data quality of which domain(s) do you primarily manage (e.g., finance, healthcare, manufacturing, or domain-agnostic)?
Challenges	- What are the biggest challenges you currently encounter in managing data quality? Have these challenges evolved over the last few years? - From your experience, what are the most common data quality challenges faced by non-DQ experts (e.g., business users)? - Which phases of DQM (definition, measurement, analysis, improvement) do you find most difficult to manage? Why? - Are there challenges related to data type (e.g., structured, unstructured, big data, real-time)? What specific issues arise when working with these types? - Are there challenges related to how your data is stored or structured (e.g., databases, data lakes, data warehouses)? What specific issues do you encounter in those environments? - Are certain DQ dimensions more difficult to manage than others (e.g., timeliness, accuracy, completeness)? What makes these dimensions particularly challenging? - Are there any other underlying factors that make DQM difficult?

Interview Protocol Section	Interview Questions
Current Approach	- How do you currently approach data quality? - What tools (if any) do you use today for DQ? - What tools have you used in the past for DQ? - In resolving DQ issues: a) Is it enough to rely on DQ expertise alone, or is business domain knowledge required? b) How often do you consult business experts or data owners? c) Are business rules predefined, or is issue resolution mostly case-by-case? d) Is communication between business and data teams a challenge?
AI in DQM	- Do you currently use AI in your DQM processes? - In what ways is AI being used in your DQM processes? - What are the reasons for not using AI in your DQM? - Where do you see AI providing the most value in DQM? - Have you applied AI to any of the specific challenges mentioned earlier? - How effective has AI been in addressing those challenges? - Why haven't you applied AI in those cases? - Should AI-empowered DQ tools be designed more for experts or for business users? Could AI reduce the need for technical expertise in DQM? - What are the biggest issues with adopting AI-empowered DQ solutions? ○ Technical complexity ○ Lack of transparency ○ Lack of explainability ○ Cost ○ Trust ○ Integration with systems ○ Regulation/compliance ○ Other (please specify)
Future	- Do you see DQM changing with the rise of technologies like GenAI? - Could GenAI improve data quality earlier in the lifecycle?

Interview Protocol Section	Interview Questions
	- Do you think GenAI will significantly change how organizations manage DQM? - Could AI play a broader role in data governance beyond DQM?

2.1.3 Participant Selection

Participants were selected using a purposive sampling strategy, which is appropriate for qualitative research aimed to gain insights from individuals with specific expertise (Palinkas et al., 2015). The selection process targeted professionals with experience in data quality management to ensure that the collected data would be relevant to the study's objectives. Participants were identified through:

- professional networks and industry contacts;
- recommendations from academic supervisor;
- authors of published research on DQM and AI solutions.

The sample was purposefully constructed to ensure diversity across multiple dimensions: participants from multidisciplinary professional backgrounds to capture a range of viewpoints; representing different countries to enhance global relevance; covering varied data-related domains to broaden contextual applicability; and reflecting a range of experience levels to provide both depth and nuance to the insights gathered. The number of interviews was not predetermined. Based on qualitative research guidelines and the need to achieve thematic saturation across the group, a target of 6–12 expert interviews was set as an initial goal (Guest et al., 2006).

2.1.4 Interview Coding

After conducting the interviews, they were transcribed using an open-source tool called *vibe.*[2] A qualitative content analysis was carried out on the transcripts using thematic coding. Inductive approach was used for labelling to allow themes to emerge from data rather than rely on predefined coding scheme (Naeem et al., 2023). The analysis involved the following steps. First, the transcripts were read through to get familiar with the content. After that, meaningful segments were identified and labelled with initial codes. These codes were then reviewed and grouped into broader themes that captured recurring ideas. The final themes were linked to the research questions and informed both the analytical framework for the results sections and the design of the tool evaluation protocol.

[2] Vibe. (2023). Vibe: Open-source desktop transcription tool. GitHub. https://github.com/thewh1teagle/vibe

2.2 Tool Analysis

This section details the methodology used to identify, select and systematically analyse AI-empowered data quality tools. It describes the search strategy for tool discovery, the selection criteria applied and the structured evaluation protocol developed to assess the tools' functionalities.

2.2.1 Tool Search Strategy

The systematic mapping review was conducted based on principles adapted from the methodology by Kitchenham et al. (2013). As part of this process, a systematic search was carried out to compile a list of AI-empowered data quality tools. Search was conducted using two primary sources: the Google search engine and the Scopus academic database. Scopus was used to identify peer-reviewed research on AI-empowered data quality tools, while Google was included to capture industry-relevant tools. Additionally, a parallel search was performed in the Web of Science database to complement the Scopus search.

The search terms were developed by combining two concept groups: (1) data quality–related terms, such as "data quality", "data profiling", "data management" and "data analysis" and (2) AI-related terms, including "artificial intelligence", "machine learning", "generative AI", "large language model", "deep learning". This combination was intended to identify tools positioned at the intersection of data quality and AI capabilities. Additional descriptors such as "software", "tool", "program" and "service" were included to narrow the results to technological solutions.

The following search query was used in Google:

(("Data quality" OR "data profiling" OR "data management" OR "data analysis") AND (software OR tool OR program OR application OR service) OR "data profiler") AND (AI OR "artificial intelligence" OR "genAI" OR "generative AI" OR "machine learning" OR ML OR "deep learning" OR "neural networks" OR "feature extraction" OR "virtual assistant" OR chatbot OR "recommending system" OR recommender OR RAG OR LLM OR "large language model")

Additionally, a literature-based search was conducted in the Scopus database to identify relevant academic publications. A parallel search in the Web of Science database was performed to complement the Scopus search. No new AI-empowered data quality tools were identified beyond those found in Scopus. Therefore, the primary analysis was based on the Scopus dataset.

The search query was limited to the article's title, abstract, and keywords (TITLE-ABS-KEY) to focus on publications where the key terms appear prominently, thereby increasing the relevance of search results and reducing irrelevant hits. Publications were restricted to the years

2015 to 2025 to reflect the period of rapid development in AI and its integration into data management technologies. Only English-language publications were included and document types were limited to articles, conference papers, chapters, short surveys, and books.

The following search query was used in Scopus:

TITLE-ABS-KEY ("Data quality tool" OR "data quality service" OR "data quality application" OR "data quality software") AND ALL("artificial intelligence" OR "machine learning" OR "generative AI" OR "large language model") AND PUBYEAR > 2014 AND PUBYEAR < 2026 AND (LIMIT-TO (DOCTYPE,"ar") OR LIMIT-TO (DOCTYPE,"cp") OR LIMIT-TO (DOCTYPE,"ch") OR LIMIT-TO (DOCTYPE,"sh") OR LIMIT-TO (DOCTYPE,"bk")) AND (LIMIT-TO (LANGUAGE,"English"))

2.2.2 Selection Criteria

To ensure that the identified AI-empowered data quality tools were relevant and suitable for analysis, a set of selection criteria was applied to both Google and Scopus search results. These criteria served as the basis for determining both inclusion and exclusion.

SC1. Tools not presented as data quality solutions (e.g., for data cleansing, anomaly detection, profiling, consistency checks) based on publicly available information (e.g., documentation, product pages or other descriptive materials) were excluded.

SC2. Tools that do not integrate AI-empowered technologies (e.g., machine learning, deep learning, generative AI) as evidenced by publicly available information (e.g. documentation, product descriptions or related materials) were excluded.

SC3. Tools that were discontinued or no longer actively maintained based on publicly available information (e.g. company websites, documentation) were excluded.

SC4. Tools that were designed exclusively for a specific domain (e.g. ontology validation) were excluded.

SC5. Tools that are not publicly accessible — either as a trial, demo, open-source repository or detailed documentation were excluded.

SC6. Articles without accessible full text were excluded from the analysis.

After applying the selection criteria to the initial list of 51 tools, 28 tools remained to be analysed further.

2.2.3 Tool Evaluation Protocol

To ensure a consistent, transparent, and comparable assessment of the selected AI-empowered data quality tools, a structured evaluation protocol was developed. This protocol was used to systematically analyse the functionality, usability, AI integration and practical fit of each tool

within the context of real-world data quality management challenges. The protocol was developed iteratively, drawing on insights gathered during expert interviews, as well as on themes derived from academic and industry literature on data quality. The interviews played a central role in identifying practitioner needs and recurring challenges, which were then reflected in the evaluation criteria.

The protocol (full protocol in Appendix II) consisted of five main sections:

- **general information** – to collect general tool details (e.g., name, website) and determine availability for analysis;
- **AI usage** – to identify AI techniques used (e.g., machine learning, NLP) and assess AI visibility, explainability, and customizability for evaluating how AI capabilities support data quality tasks;
- **functional capabilities** – to analyse core data quality functions (e.g., data profiling, anomaly detection) for assessing the extent to which tools fulfil data quality needs;
- **usability & workflow fit** – to assess usability and workflow features (e.g., user interface, collaboration) for determining how tools fit into practical workflows and user environments;
- **data processing** – to identify supported data formats (e.g., JSON, SQL, CSV, APIs) to assess integration capabilities.

Tools were evaluated using publicly available information - product websites, documentation, demo videos, user guides and free trials or open-source versions where accessible. In cases when trials were available, the tools were accessed to gain a general impression of their structure, interface and functionalities. This process enabled a systematic comparison of tools and informed the analysis presented in the results and discussion chapters.

3. Results

This section presents the key results of the study, combining qualitative insights from expert interviews with a structured analysis of AI-empowered data quality tools identified via a systematic search. Then the experts' identified needs are mapped onto the capabilities and gaps of these tools, highlighting areas of alignment and opportunities for improvement.

First, expert interviews were conducted. An interview-based thematic analysis revealed current challenges, practices and expectations surrounding data quality management and the role of AI within it. These insights directly informed the development of a multi-criteria evaluation protocol or AI-empowered DQ tools, designed to facilitate assessment whether existing tools meet real-world needs.

3.1 Interview Results: The Most Common Challenges Faced by Data Professionals in Managing Data Quality (RQ1)

A total of 35 potential interviewees were contacted, of whom 10 agreed to participate in the study. As key themes began to repeat, data collection was concluded based on emerging saturation. This approach is supported by prior research, with Guest et al. (2006) finding that thematic saturation often occurs within 6–12 interviews. This is also supported by Malterud et al. (2016), who argue that smaller samples are sufficient when participants hold high information power. Given the depth and relevance of insights collected, the sample was sufficient to meet the study's objectives.

The interviewees included ten professionals with backgrounds in data engineering, data governance, academic research and consulting. Participants were based in or had experience working across a range of countries including Estonia, Germany, Austria, Poland, the United States and Australia. Their expertise spanned domains such as finance, telecommunications, public sector services, open government data and research information systems. Several interviewees were engaged in practical data quality management activities within organizations, while others focused on academic research related to metadata quality, continuous monitoring or AI applications in data quality. Experience levels ranged from early-career professionals to senior experts with decades of experience in data-related roles. This diversity in geography, domain and role enabled the study to capture both technical and organizational aspects of DQM, as well as emerging expectations regarding AI-empowered tools.

Thematic analysis of the interview data resulted in five key themes related to data quality management and the use of AI-empowered tools: (1) Challenges in DQM, (2) Current Practices & Tools in Use, (3) Role of AI in DQM, (4) Needs & Requirements for AI Tools and (5) Opportunities for Improvement and Future Outlook. These themes emerged through iterative coding and synthesis. The findings not only provide insight into the current state and expectations surrounding data quality tooling, but also directly informed the development of the tool evaluation protocol used in this study. Selected participant quotes are included

throughout the following sections to illustrate key points. The following sections will explore each of these themes, providing a detailed overview of the findings. Selected participant quotes are included throughout the following sections to illustrate key points.

3.1.1 Theme 1: Challenges in Data Quality Management

Across the expert interviews, participants unanimously characterized data quality management as a complex and resource-intensive process, shaped by both technical and organizational constraints. Rather than isolated technical issues, participants described DQM as a context-driven, evolving practice that requires continuous negotiation between data structures, limitations of tools, internal culture and the involvement of both business and technical experts.

One of the most frequently raised challenges concerned the diverse nature of data types and the impact this has on quality monitoring. While structured data allows for more straightforward validation, it can still have its challenges.

> "Structured is the easiest, unstructured requires entirely different handling." – Participant 3
> "Each data type has its own challenges… Even structured data has unsolved issues." – Participant 2

For unstructured formats like PDFs or multimedia files, participants described not only the difficulty of accessing raw content, but also the challenge of defining what constitutes "quality" in such contexts.

"You might need OCR[3], LLMs, or manual labour to extract anything usable." – Participant 1
"We don't even have an established way to say what 'good text' is. For relational data, yes — for unstructured data like text, images, it's still unsolved." – Participant 7

In parallel, metadata emerged as a weak component in dynamic data environments. Several participants noted the challenge of keeping metadata consistent with evolving data streams, especially in real-time applications.

> "Trying to create high-quality metadata in real time is not easy." – Participant 6

> "Real-time validations are tricky – you can't run complex checks in real-time without trade-offs." – Participant 1

Several participants identified data storage structures as a hidden driver of quality issues. While relational databases were seen as more controllable, other storage systems like data lakes or data marts introduced trade-offs in quality enforcement, performance and discoverability.

[3] OCR - Optical Character Recognition

> "Relational databases enforce some quality, but people find ways around it." – Participant 2
> "Data lakes give flexibility but are messier and lead to errors that are hard to trace back." – Participant 2

Participants working with large, unstructured repositories noted the lack of contextual information and lineage within data lakes as a core challenge.

> "With data lakes, I don't know how the parts relate to each other — and without metadata, you can't trace anything." – Participant 7
> "To do serious analysis on a lake, you first need to add structure — schema, metadata, lineage. Otherwise you're flying blind." – Participant 6

Even more critical were organizational silos — isolated teams or departments collecting and storing similar data without coordination or shared access.

> "Silos mean teams don't know others are collecting similar data — massive redundancy." – Participant 5

The result is often inconsistent data schemas, duplicated efforts and missed opportunities to apply quality controls universally across an organization.

Despite the promises of modern tools, many DQM tasks are still manual or only semi-automated. Participants noted that rule-based systems, especially those relying on SQL or hard-coded logic, often lack flexibility and scale poorly.

> "Tooling is still fragmented. We've seen tools that promise a lot, but break on real use cases." – Participant 3
> "Vendors offer great demos, but try plugging into a real system… it doesn't scale." – Participant 5
> "You still need to fix the error after detection — tools often stop short of helping you solve the problem." – Participant 7

A recurring theme was the lack of organizational prioritization for DQM. While the benefits of clean data are widely acknowledged, strategic investment and long-term planning are often lacking.

> "Everyone agrees it matters, but few allocate budget or people to it." – Participant 3
> "Companies invest only when there's an audit or visible failure. Otherwise, it's low priority." – Participant 4
> "Convincing C-levels[4] that DQM is worth investing in is a full-time job in itself." – Participant 5
> "No costs arise from quality issues that remain hidden. Costs from occasional

[4] C-levels: Senior executives, such as CEOs, CIOs, and other "Chief"-titled decision-makers.

> disasters aren't associated with, and balanced against, investment in quality planning, documentation, training, deployment, review, etc." – Participant 9

Finally, several participants pointed to governance issues — particularly the absence of clear roles, fragmented responsibilities or internal restrictions on cross-team collaboration.

> "Who owns DQM? It's everyone's job — and no one's." – Participant 4
> "Departments are legally separate, so they can't collaborate easily on data fixes." – Participant 1
> "Responsibility varies wildly between teams — one person thinks they own it, another thinks someone else does." – Participant 7
> "Unless there's a champion for DQM, it becomes backlogged behind flashier projects." – Participant 5

The challenges raised by participants were used to develop the tool evaluation protocol used in this study. Frequent concerns around the complexity of different data types—especially the difficulties of managing unstructured and real-time data—highlighted the need to assess whether tools support a variety of formats beyond traditional structured inputs. Likewise, frustrations with manual rule creation, fragmented tooling and limited automation influenced the inclusion of criteria related to rule customizability and workflow integration. The recurring emphasis on the importance of domain knowledge and the human interpretability of AI systems led to a deeper focus on explainability, business user friendliness and the ability to configure or understand how quality checks are applied. Finally, observations about organizational silos and the diverse technical data environments informed the decision to evaluate tools' ability to integrate with varied storage architectures, including relational databases, data lakes and cloud services.

3.1.2 Theme 2: Current Practices & Tools in Use

Participants described a wide variety of current approaches to managing data quality in their organizations and projects, ranging from fully manual interventions to semi-automated pipelines and AI integration. Across all interviews, it was clear that tool maturity and strategy alignment vary widely across contexts, most participants described working within systems that are still fragmented and role-dependent.

Four participants described their current data quality workflows as being rule-based, with validation logic defined in SQL, Python scripts or Excel spreadsheets. In some cases, these rules were hardcoded or manually curated by data quality teams or domain experts.

> "Our data quality validations are SQL-based. We build templates and sometimes use AI to generate them, but fundamentally they're rules." – Participant 1

Others mentioned that while some automation is layered on top of these rule systems, the underlying architecture often remains fragile and rule maintenance becomes a burden as data scales or evolves. Even among more advanced users, AI is typically applied at specific points

in the workflow, such as profiling, rule suggestion or trend detection — not as a holistic solution.

Several participants expressed frustration with the lack of comprehensive tools that fit their organization's specific needs. While there are many commercial solutions available, these often require heavy customization or fail to scale effectively.

> "Most tools don't understand metadata structures. We had to customize everything." – Participant 6

How data quality is managed often depends on the role of the person involved. Data scientists and analysts may clean data locally using statistical techniques, while business users are often left out entirely or rely on workarounds.

> "Data scientists just clean data locally and move on. They rarely fix it at the source." – Participant 2
> "Researchers just clean datasets to make them usable — it's mostly manual massaging of data." – Participant 7
> "We had to write our own validation logic because the data wasn't clean enough to trust out of the box." – Participant 6

This leads to inconsistencies across teams and contributes to a disconnect between identified issues and structural resolution.

While participants agreed that business-side stakeholders often know best what "good data" looks like, their involvement in actual data quality management tasks was described as limited or inconsistent.

> "There's a disconnect. Business people spot the issue, analysts work around it, but no one goes upstream to fix it." – Participant 2
> "Business input is often needed, especially to define what needs monitoring — but they're not always looped in." – Participant 4
> "The business side doesn't always understand what's wrong — and we can't fix what they don't explain." – Participant 8
> "Many decisions get made with poor data simply because there's no feedback loop." – Participant 6

This results in a lack of true ownership over DQM processes and delays in resolving issues that cross team or system boundaries.

The reliance on rule-based systems and the need for flexible logic informed the inclusion of criteria such as rule customizability and support for SQL-based rule definition. The diversity of automation strategies, including the selective use of AI for profiling or validation generation, highlighted the importance of capturing human-AI collaboration. Meanwhile, the frustration over fragmented tooling and custom solutions reinforced the need to assess tool integrability with existing infrastructures. Finally, the limited involvement of business users informed

criteria like ease of use for non-technical users. Together, these findings helped ensure that the protocol reflects not just the capabilities of the tools, but the real-world environments in which they must operate.

3.1.3 Theme 3: Role of AI in Data Quality Management

The participants generally acknowledged the growing presence of AI in data quality workflows, although most characterized its current role as limited or experimental. Across the interviews, participants mentioned that AI was primarily used to augment specific tasks within the data quality workflow such as rule generation, anomaly detection, rather than to enable full automation. Participants also emphasized serious concerns about explainability, compliance and integration. There was agreement that while AI holds transformational potential, its use today is far from mature.

Most participants noted that AI is currently applied to narrow, well-defined tasks within broader DQM workflows. Examples included using machine learning to identify anomalies, generating validation rules based on metadata or suggesting column types based on profiling data.

> "We apply AI mainly for assessment — it's good for detecting quality issues automatically, not so much for managing them." – Participant 2

> "I use AI mainly for spotting anomalies and suggesting validation rules, but the final decision always comes back to us." - Participant 10

Newer use cases include semantic classification and prioritizing cleaning tasks, further demonstrating the incremental yet targeted use of AI.

> "We used AI to highlight semantic types — not just data types, but the meaning of fields." – Participant 6
> "We used AI to help choose which feature to clean — it prioritized attention areas." – Participant 6

Participants consistently highlighted trust, transparency and explainability as obstacles to wider AI adoption in DQM. In a domain where errors can have financial, regulatory or operational consequences, many participants viewed black-box AI systems as too risky.

> "You don't have 100% guarantee there are no hallucinations, or that the AI understands the context correctly." – Participant 1
> "For data quality projects, we always try to use more explainable models. Otherwise it's not responsible." – Participant 2
> "Business users won't trust the outputs if they don't understand how the decisions were made." – Participant 4
> "The use of ANNs[5] in DQM is extremely risky. It is purely empirical, and hence

[5] ANN - artificial neural network

a-rational... ANN-based AI techniques have no reference-point in the real world, and are not supported by any theory or model." – Participant 9

Some participants shared direct experience with compliance challenges caused by black-box models.

"We had to simplify the model — black-box models couldn't justify loan decisions for regulators." – Participant 7
"We're still not able to describe why the AI made a decision — that's a compliance issue." – Participant 7

This concern was especially pronounced in regulated sectors like finance and healthcare, where compliance and auditability are mandatory.

Rather than replacing human expertise, participants saw the ideal role of AI as collaborative — augmenting expert judgment, accelerating repetitive tasks and suggesting improvements that can be reviewed or modified by users.

"AI can generate rules, but for complex validations, experts still need to double-check them." – Participant 1
"It's not feasible to ask people for every check, but you can use AI for suggestions, not decisions." – Participant 2
"They can support us — but replacing experts entirely? Not anytime soon." – Participant 8
"You're still missing the measurement logic — AI optimizes, but doesn't know the goal unless we define it." – Participant 7

This "human-in-the-loop" approach was viewed as the most responsible and effective path forward, particularly in large organizations with a mix of technical and business stakeholders.

Several participants expressed concern that regulatory uncertainty could slow the adoption of AI-empowered DQM tools. This is especially relevant when using cloud-based models that may process sensitive data externally.

"If you're dealing with personal data and using OpenAI cloud, you might not be allowed to send it there." – Participant 1
"The AI hype is strong, but many companies are still unsure what they're even allowed to do." – Participant 5

Some suggested that local models or domain-specific fine-tuning could help mitigate these risks, but acknowledged that this approach raises cost and infrastructure complexity.

Several participants acknowledged the promise of Generative AI in data quality. However, they also emphasized that current implementations are still in early-stage exploration and most organizations have yet to establish robust GenAI workflows in DQM.

> "GenAI could be used to label metadata, manage business glossaries, or generate views for access requests." – Participant 1
> "It has high potential, but it's not going to replace the work of experts anytime soon." – Participant 2
> "We're seeing more GenAI demos than real-world implementations right now." – Participant 5

Others highlighted the potential for GenAI to support earlier-stage decision-making by replacing some of the initial expert interactions.

> "LLMs could act as oracles — instead of scheduling meetings with 3 stakeholders, ask an assistant for context." – Participant 7
> "GenAI has potential for early-stage cleaning and labelling, but we're not there yet." – Participant 8

Participants also stressed that GenAI requires strong organizational readiness, governance and human validation to be effective. These perspectives on AI directly informed how AI-related capabilities were addressed in tool analysis. Rather than treating AI as a binary feature, the protocol includes "AI Features Present", "AI Type/Technique" and "Location of AI Use". Furthermore, concerns about trust and auditability led to the inclusion of "Explainability" and "AI Customizability", helping to distinguish between tools that are merely opaque versus those that allow users to inspect or influence outputs. The repeated mention of human-AI collaboration helped define criteria focused on non-technical user access, AI-assisted rule generation and human override capabilities. Finally, references to compliance and data privacy challenges highlighted the need to track deployment mode (e.g., cloud vs. on-prem) and regulatory readiness.

3.1.4 Theme 4: Needs & Requirements for AI-empowered Data Quality Tools

While participants acknowledged the promise of AI in streamlining data quality tasks, they were also clear about what tools must offer to be genuinely usable in real-world environments. The requirements spanned technical functionality, user experience and organizational fit. Moreover, a desire for flexible, transparent and collaborative tools was consistently reflected.

A central requirement across all interviews was ease of use, particularly for non-technical users. Participants emphasized that business users often have critical data knowledge but are left out due to complexity of tools.

> "Business users might report issues, but don't have tools or access to fix them directly." – Participant 4
> "You want to enable people to contribute to data quality without requiring them to write code." – Participant 5

This included a need for simple interfaces, no-code configurations and onboarding support that would allow domain experts to participate more actively in monitoring and rule definition. Several participants also highlighted emerging interest in more conversational and intuitive interfaces, especially for enabling access to insights without needing to write queries.

> "I recently had a chat with a startup building chatbot-style reporting systems for business users — that's the right direction." – Participant 7

Participants also stressed the need for tools to handle a broad spectrum of data types and formats, including real-time, semi-structured and unstructured data. They frequently encountered tools that worked well for structured data but broke down in more complex contexts.

> "If you only work with structured data, you can use basic validations. But for PDFs or real-time streams, you need something smarter." – Participant 1
> "The more unstructured the data, the less support most tools provide." – Participant 3

Some participants noted that tools rarely consider metadata as a first-class element for quality checks, even though metadata structure and quality are increasingly crucial.

> "Metadata-aware tools are rare — most systems don't understand metadata as something to evaluate." – Participant 6

In fast-changing environments, tools also needed to adapt to evolving schemas and business requirements without requiring constant redevelopment.

For tools to succeed, they need to integrate cleanly into existing workflows. Participants described the frustration towards tools that required major architectural changes or failed to align with enterprise systems.

> "The tool has to work where the data lives — whether that's a warehouse, lake, or some API feed." – Participant 3

Participants also favoured flexible tools that support custom logic and can adapt to varied validation needs over time. In some cases, participants preferred using toolkits or libraries rather than fully integrated platforms.

> "We needed tools that worked like libraries — not full platforms, just something to plug in where we need it." – Participant 6
> "Making best practices flexible enough for future cases was essential — rigid systems didn't work." – Participant 7

They also emphasized the importance of transparency in how AI models operate and make decisions.

> "We always try to use explainable models… it's not responsible otherwise." – Participant 2
> "If I can't tell you why a record was flagged, I won't trust the tool — and neither will anyone else." – Participant 5
> "We had to go with explainable models because we couldn't justify decisions with anything else." – Participant 7

Visual outputs, logs and clear error explanations were frequently mentioned as features that build trust and facilitate adoption.

Participants' insistence on supporting diverse data environments helped justify detailed assessments of data type support, real-time compatibility and cloud-native deployment. Finally, the focus on workflow fit and integration led to evaluating tools on their API support, deployment models and ability to plug into existing pipelines.

3.1.5 Theme 5: Opportunities for Improvement & Future Outlook

Participants expressed optimism about the future of data quality management, particularly in how AI could help overcome existing limitations. However, vision for the future was not one of fully autonomous systems. Instead, it centred on building context-aware, human-aligned and adaptable frameworks that support collaboration, scale with complexity and respect governance requirements. Across the interviews, there was a call for more intelligent validation, better integration of domain knowledge and a stronger focus on tools that fit organizational realities.

Many participants envisioned future tools offering more context-aware quality checks - systems that could dynamically adapt to the content, structure and purpose of data. Rather than hard-coded validations, the goal was to use machine learning or rules engines that could infer what "quality" means in a given situation.

> "You could have AI agents that apply predefined checks but also surface new insights from the data itself." – Participant 1
> "Tools should understand when something looks wrong, not just when it's missing." – Participant 5
> "Smart validation systems should know when something isn't wrong but just unusual." – Participant 6

Participants also emphasized the potential of proactive detection, where tools spot issues before data enters critical systems. This means not only flagging anomalies, but also suggesting remediation strategies.

A recurring idea was the importance of designing systems where humans remain part of the loop, not just to validate AI suggestions, but to enrich them with business logic and evolving knowledge. Participants argued for tools that offer both automation and interactive touchpoints where users can influence logic, tune parameters or override results.

> "You want automation, yes, but you also want checkpoints where someone can say: this doesn't make sense here." – Participant 4
> "Tools should learn from the organization — from how people actually use the data." – Participant 3
> "If you remove people too early, you lose nuance. DQM is full of edge cases." – Participant 7

The goal is not to reduce headcount but to reallocate expertise more effectively, giving people time to focus on high-value work rather than manual data correction.

Participants also pointed to the need for tools that integrate natively with existing data platforms, reducing the overhead of adoption and enabling smoother workflows. Rather than replacing what exists, future tools should embed themselves within the fabric of modern data ecosystems.

> – Participant 3
> "The tool should be invisible when it's working right — it should feel like part of your environment." – Participant 5
> "A validation rule shouldn't be a report — it should be a transformation that fits into the pipeline." – Participant 6

Improved interoperability was seen as key not only to adoption but to organizational trust in DQM tooling.

Several participants stressed the opportunity to blend AI capabilities with "human-authored" rules and domain knowledge. In this hybrid vision, tools would draw from both historical data patterns and encoded business logic.

> "AI doesn't know your business. But it can help once it's told what matters." – Participant 4
> "The best systems will combine what the company knows with what the data shows." – Participant 1
> "Business logic encoded in workflows — that's where AI should plug in, not replace them." – Participant 7

Finally, participants spoke about the importance of organizational culture and leadership buy-in for enabling meaningful improvements. Better tools are essential, but without strategic alignment, ownership and accountability, even the best technology will fall short.

> "We need to show management real cost numbers of bad data — that's when they start caring." – Participant 1
> "Unless there's a champion, DQM becomes a side project. We need it embedded in the way people work." – Participant 5
> "Tooling without accountability is shelfware. People need to see themselves in the system." – Participant 8

This echoed the idea that tools alone aren't the solution. The future of DQM also depends on fostering a strong data culture encompassing mindsets and incentives.

Participants' emphasis on seamless integration into existing environments supported the inclusion of "Workflow Integration", "Cloud Compatibility" and "Collaboration Features." Their call for context-aware and human-aligned tooling shaped criteria such as "Human-AI Collaboration", "Custom Rule Flexibility" and "Designed for Business Users." Additionally, the importance of cultural alignment and shared accountability was reflected in the evaluation of "Ease of Onboarding" and "Team Permissions".

By grounding the evaluation protocol in both current practices and expert visions of future DQM, it aims to assess not just technical capabilities, but whether a tool is realistically deployable and aligned with how data work is done across teams and organizations. Rather than evaluating tools on theoretical capabilities alone, the protocol was designed to reflect what practitioners genuinely need to adopt and succeed with AI-empowered DQM tools in practice.

3.2 Analysis of AI-empowered Data Quality Tools (RQ2)

This section presents the search results, selection process and analysis of AI-empowered data quality tools. It assesses their capabilities against the practical challenges identified in expert interviews.

3.2.1 Tool Search Results

Google search returned 14 results, including industry reports and company websites listing AI-empowered data quality tools. Table 2 shows the results that led to 26 tools included in the analysis. The full list of search results can be found in the Appendix III.

Table 2. Overview of Tools Identified through Google Search.

Reference	Tools Mentioned in Source
(Jackson, 2024)	RapidMiner Tableau Qlik Polymer Databricks Unified Data Analytics Platform Sisense The KNIME Analytics Platform IBM Watson Analytics Google Cloud Smart Analytics Microsoft Azure Machine Learning
(*SAS*, n.d.)	SAS

Reference	Tools Mentioned in Source
(*Software*, n.d.)	HPE
(*Julius AI | Your AI Data Analyst*, n.d.)	Julius AI
(*Informatica Data Quality and Observability*, n.d.)	Informatica
(*Welcome - YData Profiling*, n.d.)	YData Profiling
(*Effortless Data Quality, Infused with AI*, n.d.)	Ataccama
(*Quality Monitoring & Data Profiling Tool Development ☑ Acropolium's Case Study*, n.d.)	Acropolium
(*Digna - AI-Powered Data Quality for Data Warehouses & Co., Made in Europe*, n.d.)	Digna AI
(AnalytixLabs, 2025)	RapidMiner Talend ThoughtSpot KNIME Google Sheets DataRobot Akkio IBM Watson Analytics H20.ai Microsoft Power BI Tableau Luzmo

Scopus search resulted in 32 papers. After manual review, 25 additional tools were identified. Table 3 includes only those papers that led to identified tools; the full list of papers is provided in the Appendix IV.

Table 3. Overview of Tools Identified Through Scopus search.

Reference	Tools Mentioned in Source
(Ustunboyacioglu, Kumara, Di Nucci, Tamburri, & Van Den Heuvel, 2024)	(Py)Deequ Pandera Data Build Tool (DBT) Great Expectations (GX) TensorFlow Data Validation (TFDV)
(Zhou et al., 2024)	Kylo MobyDQ Apache Griffin SQL Power Architect Aggregate Profiler YData Quality DataCleaner WinPure SQL Power DQguru Deequ Dataedo OpenRefine Great Expectations Soda Ataccama ONE whylogs Evidently
(Ustunboyacioglu, Kumara, Di Nucci, Tamburri, & van den Heuvel, 2024)	PyDeequ AWS
(Fadlallah et al., 2023)	QuaIIe SparkDQ
(Ehrlinger et al., 2023)	DQ-MeeRKat HoloDetect
(Azeroual & Lewoniewski, 2020)	DataCleaner
(Brennan, 2017)	Dacura Quality Service RDFUnit
(Božić et al., 2016)	Dacura Quality Service

3.2.2 Tool Selection

An initial list of tools was compiled through a systematic Google and Scopus search, resulting in a total of 51 candidate tools. Based on the selection criteria described in the methodology section, 23 tools were excluded from the analysis.

In total, 28 tools met being included in the further analysis. A list of the analysed tools, along with inclusion status and specific exclusion criteria, is presented in Table 4 below.

Table 4. Overview of candidate tools, inclusion status and exclusion criteria

Tool Name	Website	Included	Exclusion Criteria
RapidMiner	https://altair.com/altair-rapidminer	Yes	
Tableau	https://www.tableau.com	Yes	
Qlik	https://www.qlik.com/us	Yes	
Polymer	https://www.polymersearch.com	Yes	
DataBricks	https://www.databricks.com	Yes	
Sisense	https://www.sisense.com	Yes	
KNIME	https://www.knime.com	Yes	
IBM Watson Analytics	https://www.ibm.com/products/watson-studio	Yes	
Google Cloud Smart Analytics	https://cloud.google.com	Yes	
Microsoft Azure Machine Learning	https://azure.microsoft.com/	Yes	
SAS Vija	https://www.sas.com/	Yes	
HPE Machine Learning Data Management	https://www.hpe.com/au/en/software.html	No	SC2
Julius AI	https://julius.ai	Yes	
Informatica	https://www.informatica.com	Yes	
YData Profiling	https://docs.profiling.ydata.ai/latest/	No	SC2
Ataccama	https://www.ataccama.com	Yes	
Acropolium	https://acropolium.com/	No	
Digna AI	https://www.digna.ai	No	SC5
Talend	www.talend.com	Yes	
ThoughtSpot	https://www.thoughtspot.com	Yes	
Google Sheets	https://sheets.google.com/	No	SC1
DataRobot	https://www.datarobot.com	Yes	
Akkio	https://www.akkio.com	Yes	
H20.ai	https://h2o.ai	Yes	
Microsoft Power BI	https://powerbi.microsoft.com	Yes	
Luzmo	https://www.luzmo.com	Yes	

Tool Name	Website	Included	Exclusion Criteria
Kylo	https://kylo.io	No	SC3
MobyDQ	https://ubisoft.github.io/mobydq/	No	SC2
Apache Griffin	https://griffin.apache.org	No	SC2
SQL Power Architect	https://bestofbi.com/products/sql-power-architect-data-modeling/	No	SC2
Aggregate Profiler	https://sourceforge.net/projects/dataquality/	No	SC2, SC5
DataCleaner	https://datacleaner.github.io	No	SC2
WinPure	https://winpure.com	Yes	
SQLPowerDQGuru	https://bestofbi.com/products/sql-power-dqguru-data-quality/	No	SC2
Dataedo	https://dataedo.com	Yes	
Openrefine	https://openrefine.org	No	SC2
Soda	https://soda.io	Yes	
whylogs	https://whylabs.ai	No	SC2
Evidently AI	https://www.evidentlyai.com	Yes	
(Py)Deequ	https://github.com/awslabs/deequ	No	SC2
Pandera	https://www.union.ai/pandera	No	SC2
Data Build Tool (dbt)	https://www.getdbt.com	Yes	
Great Expectations (GX)	https://greatexpectations.io	No	SC2
Tensorflow Data Validation (TFDV)	https://www.tensorflow.org	Yes	
QuaIIe	https://github.com/zbw/qualle	No	
SparkDQ	https://github.com/PasaLab/SparkDQ	No	SC2, SC3
DQ-MeeRKat	https://github.com/lisehr/dq-meerkat	No	SC2
HoloDetect	https://arxiv.org/abs/1904.02285	Yes	
Dacura Quality Service	-	No	
RDFUnit	https://github.com/AKSW/RDFUnit	No	SC2, SC4
Oracle Enterprise Data Quality (EDQ)	https://www.oracle.com/	No	SC2

The tools were assessed using an evaluation protocol developed through literature synthesis and expert interviews (Section 3.1), consisting of 5 sections and 45 criteria, covering:

- General Information (e.g. website, documentation, trialability)
- AI Usage (e.g. AI type, AI explainability)
- Functional Capabilities (e.g. match detection, anomaly detection)
- Usability & Workflow Fit (e.g. ease of onboarding, collaboration features)

- Data Processing (e.g. API support, relational database support)

Each tool was systematically reviewed based on documentation, product pages, demo environments and free trial access where possible. The following section presents the results of this evaluation, organized thematically by protocol criteria.

3.2.3 Tool Evaluation Results

This section presents a summary of the evaluation findings across the selected AI-empowered data quality tools, highlighting common patterns, capabilities and gaps.

3.2.3.1 AI Usage

First, AI usage was analysed. The results show that hybrid approaches that combine multiple AI techniques were the dominant configuration among the evaluated tools. Out of 28 tools, 27 (96.4%) featured combinations of machine learning, natural language processing, generative AI or deep learning. Single-method implementations were rare, only 1 tool (HoloDetect) relied exclusively on machine learning.

With regard to visibility and configurability, 20 tools (71.4%) integrated AI features that were both user-visible and user-configurable. Another 7 tools (25%) presented AI results directly to the user but offered no adjustable parameters. Two tools (7.1%) combined visibility and configurability through separate modules. Three tools (10.7%) provided partial or limited configurability.

Explainability was most commonly delivered through visualizations, included in 11 tools (39.2%). More extensive mechanisms combining white-box modelling, visuals and logs were present in 10 tools (35.7%), and 4 tools (14.3%) offered white-box outputs with visual support. Three tools (10.7%), RapidMiner, DataBricks and KNIME, limited explainability to specific models. While nearly all tools included some form of explainability, its depth and formality varied considerably.

Customizability of AI models was limited across the sample. Eleven tools (39.3%) supported both training and rule extension. Four tools (14.3%) offered trainable models without rule logic customization. Seven tools (25%) relied solely on pre-trained components, while three others (10.7%) provided limited customization through AutoML or prompt-level interaction.

These results indicate that while AI usage is broadly present, advanced functionality such as explainability and customizability remains uneven. Most tools have limited alignment with expert-identified needs in operational data quality management.

3.2.3.2 Functional Capabilities

This section examines the core data quality functionalities supported by the analysed tools.

Data profiling was the most consistently supported functionality. Among the 28 tools, 27 (96.4%) support data profiling features such as value distribution analysis, null detection or data type recognition - either natively or via extensions or code-based modules. Only one tool, HoloDetect, did not offer standard data profiling functionality. In the case of dbt, profiling was supported through a community-maintained package (dbt-profiler), while TensorFlow Data Validation (TFDV) includes profiling as a core capability but required Python-based implementation. Anomaly detection was supported by 24 tools (85.7%), making it the second most common capability. Error reporting was also prevalent, present in 22 tools (78.6%), typically in the form of validation failure logs, issue tagging or exception highlighting.

Custom rule definition was supported by 19 tools (67.9%), enabling users to define validation logic through either code, visual builders or configuration panels. Rule-based data quality checks were implemented in 18 tools (64.3%), and 16 tools (57.1%) allowed users to define rules using SQL or similar query languages. However, predefined data quality dimensions, such as completeness, consistency, or accuracy, were included in only 9 tools (32.1%). Similarly, only 11 tools (39.2%) provided a reusable rules repository, indicating limited support for operationalizing rules across workflows or datasets.

Dashboards for monitoring data quality were offered by 17 tools (60.7%). These typically included issue trend visualizations, validation pass/fail rates and filtering options by dimension or rule. Match detection features, used in deduplication or entity resolution, were present in 12 tools (42.9%). These implementations varied from fuzzy matching algorithms to manual matching workflows.

Data cleansing operations—such as standardization, trimming or null imputation—were supported by 21 tools (75%). Data enrichment functionality, including the addition of reference data or derived values, was available directly in 17 tools (60.7%) and through joins can be done in two(7.1%) tools. Master data management (MDM) features were the least common, present in only 6 tools (21.4%), and typically limited to basic golden record creation or entity linking.

The results indicate that most tools provide strong baseline support for profiling, anomaly detection, and error reporting. However, support for rule reusability, semantic frameworks, and master data governance remains limited. While many platforms allow users to define and execute rules, fewer provide the infrastructure needed to scale those practices across teams or evolving data environments.

3.2.3.3 Usability and Workflow Fit

Usability and workflow fit are critical for the adoption and sustained use of data quality tools, especially in organizations with diverse user groups and varying levels of technical expertise.

This section evaluates the tools' onboarding support, user interface design, business user accessibility, integration into data workflows and collaborative features.

A total of 22 tools (78.6%) provided modern graphical user interfaces, including web-based dashboards, no-code editors and visual rule-building environments. Three tools (10.7%), TensorFlow Data Validation, dbt, and HoloDetect, were code- or CLI-based, making them more suitable for technical users familiar with scripting environments. The remaining four tools (14.3%) offered a hybrid experience, combining visual dashboards with configuration via code or scripts.

Business user-friendliness was observed in 13 tools (46.4%), as evidenced by documentation, tutorials or support materials tailored to non-technical users. These platforms typically offered simplified interfaces, natural language interaction or no-code configuration to support non-technical roles. Among them, tools such as Tableau, Power BI and Julius AI, provided intuitive dashboards and user-guided exploration features. 12 tools (42.9%) were classified ad partially designed for business users, meaning that some level of technical knowledge or coding is needed for using them. The remaining 3 tools (10.7%) either lacked such support or required substantial technical skills, making them less suitable for inclusive, cross-functional teams.

Each system provided structured documentation, tutorials, support portals or example configurations. While the level of interactivity varied - from status documentation to guided walkthroughs - the presence of learning materials was consistent. This suggests that vendors recognize the importance of user enablement regardless of technical background.

Workflow integration was universally present but implemented through different mechanisms:

- 14 tools (50%) combined API access with GUI-based workflows, allowing for both visual configuration and programmable automation;
- 6 tools (14.3%) supported hybrid integration, combining APIs, GUIs and no-code pipelines (e.g., Google Cloud Smart Analytics, Informatica);
- 5 tools (17.9%) relied on API-only integration;
- 3 tools were code- or CLI-based: dbt (CLI-based pipeline management), TensorFlow Data Validation (Python library for data validation), HoloDetect (a research prototype requiring code-based integration).

This breakdown shows a clear preference for hybrid API-GUI models but also underscores that no-code accessibility remains limited. Tools requiring CLI or code-based integration may be powerful but are less suitable for low-code environments or teams without strong engineering capacity.

Collaboration features were supported in 26 tools (92.9%). These included team-based permission models, shared rule libraries, in-app commenting or collaborative dashboards. Collaboration was often enabled through enterprise deployment features (e.g., KNIME Server, Power BI workspaces) or via cloud-based platforms.

3.2.3.4 Data Processing Capabilities

This section examines the types of data formats, storage systems and integration methods supported by the evaluated tools. Support was coded based on explicit mention of each format or connection type in product documentation or configuration options.

Delimited file formats were widely supported. Specifically, 27 tools (96.4%) supported CSV or TSV files, and 24 tools (85.7%) supported spreadsheets. JSON compatibility, important for semi-structured data handling, was confirmed in 27 tools (96.4%).

Connectivity to structured data systems was similarly strong. Relational databases were supported by 26 tools (92.9%), and data warehouses by 24 tools (85.7%). Support for NoSQL databases was less common, present in 21 tools (75%). Although most platforms integrate well with tabular sources, support for document-based or schema-less data remains incomplete.

Data lakes were supported by 22 tools (78.6%). This includes direct access to object storage platforms such as Amazon S3 or HDFS, or native compatibility with lakehouse architectures. The lower coverage rate suggests that although data lakes are increasingly central to enterprise architecture, many tools have not fully adapted to their technical requirements.

API-based data exchange was supported in 23 tools (82.1%), enabling ingestion from streaming systems, service layers, or pipeline orchestration tools. Cloud-native deployment support was reported in 25 tools (89.3%). This includes containerized deployment, SaaS availability or optimized hosting within cloud infrastructure.

The data indicate that while most tools support traditional file formats, relational systems, and cloud-based deployment, compatibility with NoSQL databases and data lakes is still limited. As modern architectures increasingly rely on semi-structured and large-volume storage systems, these gaps may constrain the applicability of some tools in contemporary data environments.

3.2.4 Recommendations for AI-empowered DQM Tools

The interviews conducted for this study revealed a consistent set of expectations, frustrations, and aspirations regarding the role of tools in data quality management. Rather than calling for fully automated systems, experts emphasized the need for context-aware, human-aligned tools that support trust, adaptability, and collaborative governance. This includes keeping humans in the loop, not only to oversee AI-assisted processes, but also to ensure transparency and explainability in how decisions are made. The following directions reflect these priorities and highlight areas where current solutions remain underdeveloped.

1.Transparency and Explainability as Key Considerations

Explainability was cited as a prerequisite for trust in AI-enabled validation. Participants in regulated industries, such as healthcare and finance, specifically warned against opaque outputs, noting that *"you can't justify decisions if you don't understand how they were made"*

(Participant 7). This concern extended to business users as well, who were described as reluctant to act on results without clear justifications. As reflected in Theme 3, systems that aim to influence operational data quality must offer visual explanations, traceable logic or interpretable model diagnostics.

2. Systems Should Support Contextual and Domain-Aware Adaptation

Participants emphasized that general-purpose validation logic rarely aligns with domain-specific realities. Several noted that *"pretrained models don't know your business"* and advocated for the ability to fine-tune models or rules to local data structures and semantics. While some tools enable rule editing, very few support adaptive learning. As captured in Theme 4, real-world deployment requires not just automation but alignment with organizational context. For example, the use of LLMs offers potential avenues for better contextual adaptation. While LLMs have limitations in deeply understanding context, providing structured and relevant inputs may still enable them to support more targeted decisions and generate outputs that reflect the business logic of the entity in question. This approach opens opportunities for aligning model behaviour more closely with domain-specific requirements.

3. Human-AI Collaboration over Full Automation

Rather than replacing experts, participants favoured systems that augment their workflows. AI was seen as effective for surfacing suggestions or anomalies but not for making final decisions. *"AI can suggest, but humans decide,"* as summarized by Participant 2. Tools must reflect this collaborative ideal by enabling user-in-the-loop correction, especially in sensitive domains.

4. Metadata and Lineage Should Be Treated as Integral Components

Across interviews, metadata quality emerged as both essential and under-supported. Experts described metadata as *"fragile, especially in real-time systems"* and warned that poor lineage tracking renders validation meaningless. While some tools include passive metadata viewers, few enable active monitoring, validation or recovery. Tools that treat metadata as a primary validation object would better support evolving data architectures. Emerging AI and GenAI capabilities can help address this gap by inspecting metadata for anomalies, enriching missing fields and generating lineage or glossary entries. This enriched metadata can then inform more accurate and context-aware data quality management.

5. Validation Must Extend Beyond Detection Toward Remediation

Experts criticized current tools for stopping at problem identification. This aligns primarily with the "Measure" phase of the TDQM framework. While some tools help define and detect issues, few offer meaningful support for the later TDQM phases: analysing root causes or implementing improvements. *"You still need to fix the error after detection—tools often stop short of helping you solve the problem"* (Participant 7). Future systems must support not just anomaly detection but actionable remediation strategies: rule suggestions, enrichment pathways or automated repair options guided by human oversight. As emphasized in Theme 5, proactive and prescriptive capabilities are essential. For example, systems that suggest likely causes of detected anomalies, recommend context-aware rule corrections or adapt based on

how users previously resolved similar issues.

6. Systems Should Support Unstructured and Streaming Data

Despite the growing volume of unstructured data, most tools remain built for static, structured inputs. Several participants described the need to manually extract relevant content and convert files to compatible formats, before quality validation could begin. *"For unstructured data like text or images, it's still unsolved",* as noted in Theme 1. Effective solutions must either embed preprocessing (e.g., OCR, semantic parsing) or offer pluggable modules to reduce this bottleneck.

7.Generative AI Should Be Applied with Caution and Clarity

While participants were optimistic about GenAI's future role, they consistently noted that current applications are experimental. *"We're seeing more GenAI demos than real-world implementations,"* said Participant 5. To be useful, such features must be explainable to avoid hallucinations or compliance risks. GenAI integration should prioritize clarity over novelty (Themes 3 and 5).

The tool evaluation confirmed that AI is increasingly integrated into data quality platforms. However, significant gaps remain in areas such as explainability, unstructured data support and collaborative workflows. By systematically aligning tool capabilities with expert-identified needs, this study surfaces not only areas of maturity but also critical limitations in current solutions. The resulting analysis and conceptual mapping provide a foundation for future design efforts and help bridge the disconnect between tool functionality and real-world data quality challenges.

4. Discussion

This study set out to explore how artificial intelligence is shaping the field of data quality management. The findings show that while many tools now include AI features, there is still a mismatch between what professionals need and what current solutions deliver.

From the interviews, several recurring challenges emerged. Data professionals described data quality management as complex and context-dependent. They noted that poor support for unstructured and real-time data often limits what can be validated in practice. Metadata handling, lineage tracking and schema changes were also identified as pain points. These challenges are not only technical, but also organizational. Silos between teams, unclear ownership and weak incentives all contribute to inconsistent quality efforts. This gives a grounded picture of the kinds of problems professionals face on a daily basis.

As Participant 7 emphasized, "DQM is full of edge cases", highlighting the complexity and need for handling data quality issues. The organizational culture and leadership buy-in were also repeatedly cited as critical enablers for meaningful DQM improvements. Without strategic alignment and clear accountability, even the most sophisticated tools risk becoming underutilized or ineffective.

The evaluation of 28 AI-powered tools helped identify what kinds of solutions are available today. Most tools offer strong support for structured data profiling and rule-based checks. Many include dashboards, alerts and anomaly detection modules. Some platforms go further, using machine learning for rule suggestions or semantic classification. However, few tools move beyond detection to support actual remediation or proactive action. Generative AI features exist, but are limited to interface enhancements or metadata suggestions.

When comparing tool capabilities with the challenges raised in the interviews, a gap becomes clear. Few tools offer transparency into how their AI works. Explainability remains limited, even though professionals emphasized its importance for trust and compliance. Business users are often excluded due to lack of technical knowledge. While the tools provide valuable functions, they rarely reflect the workflows, roles or integration needs described by experts. Notably, Participant 5 captured a key need by stating, "You want to enable people to contribute to data quality without requiring them to write code", underscoring the importance of no-code and low-code solutions that can empower non-technical users. This human-AI collaboration is essential to making data quality management accessible across roles and expertise levels.

This suggests that the AI-empowered DQM field is still developing. Even among mature tools, many fall short of real-world expectations. The potential is there, but it has not yet translated into full alignment with professional needs. This is especially visible in cases where tools claim automation but fail to adapt to local context or involve human decision-makers in meaningful ways.

Additionally, the findings align with existing literature that stresses the importance of reallocating human expertise towards higher-value tasks, rather than replacing it. As noted in

prior studies (e.g., Batini et al., 2009; Zhang et al., 2023), effective DQM requires keeping humans in the loop, with AI augmenting rather than substituting expert judgment.

Some limitations should be acknowledged. This study relied on publicly available information for tool analysis. While hands-on testing would have provided additional insights, not all tools offered trial environments or demo access. To ensure a consistent and systematic comparison across all tools, a document-based approach was chosen. However, such an approach also presents potential risks: vendor documentation may omit certain limitations, overstate capabilities that are not yet fully implemented or be outdated. Additionally, the understanding of AI capabilities may be either too broad, reflecting generic marketing claims, or too narrow, missing important implementation details.

The sample of expert interviews was limited to ten professionals, which may be considered small by some sources. However, this study followed the qualitative research principle that data collection can be concluded once thematic saturation is reached. Based on this criterion, the sample was deemed sufficient. Still, expanding the interview base in future research could help deepen understanding in specific areas, such as sector-specific requirements, organizational data maturity levels or regional differences in DQM practices.

Overall, the combination of interviews and tool analysis provides a strong basis for understanding both what data professionals need and what the current tools offer.. Future research and development efforts should prioritize the alignment between expert-identified needs and tool capabilities, focusing on areas where progress has been made and where significant gaps persist.

Conclusion

This thesis studied the role of artificial intelligence in data quality management. The goal of this study was to explore whether current AI-empowered tools address the real-world challenges faced by data quality professionals. This was approached through expert interviews to identify key needs, followed by a structured analysis of tools based on those insights.

The interviews revealed many ongoing challenges in DQM. These included poor support for unstructured data, a lack of explainability and weak collaboration features. Experts want tools that are more transparent, more flexible and easier to use across teams.

The analysis of identified 28 AI-powered DQ tools confirmed these needs are not fully met. While most tools offer strong profiling and rule definition features, few go beyond detection. Many stop short of supporting fixes, trust-building or advanced data types. While generative AI features are starting to appear in DQ tools, their adoption is still in its infancy and largely experimental.

Together, the findings show that the field of AI in DQM is still growing. Tools are becoming more powerful, but they often fall short of real-world expectations.

The key contributions of this study include (1) an expert-informed evaluation protocol to assess AI-empowered DQM tools and (2) an overview of how selected tools perform when assessed against this protocol. The work helps tool developers understand where to focus next and offers researchers a structured basis for future studies. It also offers practical recommendations for improving explainability, collaboration and support for complex data types.

In the future, researchers could test these tools in live settings and include more users in the evaluation. As AI becomes more integrated into data platforms, it will be important to track how these systems evolve and whether they begin to meet the high expectations of data professionals.

References

1. Abdullah, N., Ismail, S. A., Sophiayati, S., & Sam, S. M. (n.d.). *Data Quality in Big Data: A Review*.
2. Abraham, R., Schneider, J., & Vom Brocke, J. (2019). Data governance: A conceptual framework, structured review, and research agenda. *International Journal of Information Management*, *49*, 424–438. https://doi.org/10.1016/j.ijinfomgt.2019.07.008
3. Altendeitering, M., Dübler, S., & Guggenberger, T. (2022). *Data Quality in Data Ecosystems: Towards a Design Theory*. 28th Americas Conference on Information Systems, AMCIS 2022. Scopus.
4. Altendeitering, M., & Guggenberger, T. M. (2024). *Data Quality Tools: Towards a Software Reference Architecture*. 6159–6168. Scopus.
5. Altendeitering, M., Guggenberger, T. M., & Möller, F. (2024). A design theory for data quality tools in data ecosystems: Findings from three industry cases. *Data & Knowledge Engineering*, *153*, 102333. https://doi.org/10.1016/j.datak.2024.102333
6. Altendeitering, M., Pampus, J., Larrinaga, F., Legaristi, J., & Howar, F. (2022). Data sovereignty for AI pipelines: Lessons learned from an industrial project at Mondragon corporation. *Proceedings of the 1st International Conference on AI Engineering: Software Engineering for AI*, 193–204. https://doi.org/10.1145/3522664.3528593
7. AnalytixLabs. (2025, March 7). Top 12 AI Tools for Data Analysis To Include In Your Tech Stack. *Medium*. https://medium.com/@byanalytixlabs/top-12-ai-tools-for-data-analysis-to-include-in-your-tech-stack-be217a762762
8. Angelov, P. P., Soares, E. A., Jiang, R., Arnold, N. I., & Atkinson, P. M. (2021). Explainable artificial intelligence: An analytical review. *WIREs Data Mining and Knowledge Discovery*, *11*(5), e1424. https://doi.org/10.1002/widm.1424
9. Ashofteh, A., & Bravo, J. M. (2021). Data science training for official statistics: A new scientific paradigm of information and knowledge development in national statistical systems. *Statistical Journal of the IAOS*, *37*(3), 771–789. https://doi.org/10.3233/SJI-210841
10. Azeroual, O. (2024). Can generative AI transform data quality? A critical discussion of ChatGPT's capabilities. *Academia Engineering*, *1*(4). https://www.academia.edu/2994-7065/1/4/10.20935/AcadEng7407
11. Azeroual, O., & Lewoniewski, W. (2020). How to Inspect and Measure Data Quality about Scientific Publications: Use Case of Wikipedia and CRIS Databases. *Algorithms*, *13*(5), 107. https://doi.org/10.3390/a13050107
12. Bak, M., Madai, V. I., Fritzsche, M.-C., Mayrhofer, M. T., & McLennan, S. (2022). You Can't Have AI Both Ways: Balancing Health Data Privacy and Access Fairly. *Frontiers in Genetics*, *13*. https://doi.org/10.3389/fgene.2022.929453
13. Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for Data Quality Assessment and Improvement. *ACM Comput. Surv.*, *41*. https://doi.org/10.1145/1541880.1541883
14. Bauskar, S. (2024). *An Predictive Analytics Or Data Quality Assessment Through Artificial Intelligence Techniques*. https://doi.org/10.2139/ssrn.4980802
15. Božić, B., Brennan, R., Feeney, K. C., & Mendel-Gleason, G. (2016). *Describing reasoning results with RVO, the reasoning violations ontology*. *1585*, 62–69. Scopus.
16. Brennan, R. (2017). Challenges for Value-driven Semantic Data Quality Management: *Proceedings of the 19th International Conference on Enterprise Information Systems*, 385–392. https://doi.org/10.5220/0006387803850392
17. Buiten, M. C. (2019). Towards Intelligent Regulation of Artificial Intelligence. *European Journal of Risk Regulation*, *10*(1), 41–59. https://doi.org/10.1017/err.2019.8
18. Cheng, W.-C., & Chiu, M.-H. P. (2023). Factors Influencing Practitioners' Experience of Utilizing Open Government Data. *Journal of Library and Information Studies*, *21*(2), 119–151. Scopus. https://doi.org/10.6182/jlis.202312_21(2).119
19. Chuprov, S., Viksnin, I., Kim, I., Melnikov, T., Reznik, L., & Khokhlov, I. (2021). *Improving knowledge based detection of soft attacks against autonomous vehicles with reputation, trust*

and data quality service models. 115–120. Scopus. https://doi.org/10.1109/SMDS53860.2021.00025
20. Cichy, C., & Rass, S. (2019). An Overview of Data Quality Frameworks. *IEEE Access*, *7*, 24634–24648. https://doi.org/10.1109/ACCESS.2019.2899751
21. Daly, M., Kelleher, J., & Brennan, L. (2025). The role of trust and attitudes in organizational adoption of artificial intelligence: A longitudinal study. *Technological Forecasting and Social Change*, 191, Article 122488. https://doi.org/10.1016/j.techfore.2025.122488
22. *Data Quality: Why It Matters and How to Achieve It*. (n.d.). Gartner. Retrieved May 2, 2025, from https://www.gartner.com/en/data-analytics/topics/data-quality
23. De Haan, M., Van der Vlist, F., & Van der Meer, A. (2024). Challenges and opportunities in managing unstructured data: A systematic literature review. *Information Processing & Management*, *61*(2), 103571. https://doi.org/10.1016/j.ipm.2023.103571
24. Dhoni, P. (2023). *Enhancing Data Quality through Generative AI: An Empirical Study with Data*. https://doi.org/10.36227/techrxiv.24470032.v1
25. DiCicco-Bloom, B., & Crabtree, B. (2006). The qualitative research interview. *Medical Education*, *40*, 314–321. https://doi.org/10.1111/j.1365-2929.2006.02418.x
26. *digna—AI-Powered Data Quality for Data Warehouses & Co., Made in Europe*. (n.d.). Digna Website. Retrieved May 1, 2025, from https://www.digna.ai/
27. *Effortless data quality, infused with AI*. (n.d.). Retrieved May 1, 2025, from https://www.ataccama.com/platform/data-quality
28. Ehrlinger, L., Gindlhumer, A., Huber, L.-M., & Wöß, W. (2021). DQ-MeeRKat: Automating Data Quality Monitoring with a Reference-Data-Profile-Annotated Knowledge Graph: *Proceedings of the 10th International Conference on Data Science, Technology and Applications*, 215–222. https://doi.org/10.5220/0010546202150222
29. Ehrlinger, L., Werth, B., & Wöß, W. (2023). Automating Data Quality Monitoring with Reference Data Profiles. *Communications in Computer and Information Science*, *1860 CCIS*, 24–44. Scopus. https://doi.org/10.1007/978-3-031-37890-4_2
30. Ehrlinger, L., & Wöß, W. (2022). A Survey of Data Quality Measurement and Monitoring Tools. *Frontiers in Big Data*, *5*, 850611. https://doi.org/10.3389/fdata.2022.850611
31. Fadlallah, H., Kilany, R., Dhayne, H., El Haddad, R., Haque, R., Taher, Y., & Jaber, A. (2023). Context-aware Big Data Quality Assessment: A Scoping Review. *Journal of Data and Information Quality*, *15*(3), 1–33. https://doi.org/10.1145/3603707
32. Gami, S., Remala, R., & Mudunuru, K. R. (2024). AI-Driven Adaptive Data Cleansing: Automating Error Detection and Correction for Dynamic Datasets. *International Journal of Computer Trends and Technology*, *72*, 159–164. https://doi.org/10.14445/22312803/IJCTT-V72I11P117
33. Guest, G., Bunce, A., & Johnson, L. (2006). How many interviews are enough? An experiment with data saturation and variability. Field Methods, 18(1), 59–82.
34. Guo, C., & Jiao, P. (2024). Evaluation and Optimization of Machine Learning Algorithms in Personalized Marketing. *2024 Third International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE)*, 1–5. https://doi.org/10.1109/ICDCECE60827.2024.10549387
35. Hardinges, J., Simperl, E., & Shadbolt, N. (2024). We Must Fix the Lack of Transparency Around the Data Used to Train Foundation Models. *Harvard Data Science Review*, *Special Issue 5*. https://doi.org/10.1162/99608f92.a50ec6e6
36. Harrington, J. L. (2016). *Relational Database Design and Implementation*. Morgan Kaufmann.
37. Haruki, Y., Kato, K., Enami, Y., Takeuchi, H., Kazuno, D., Yamada, K., & Hayashi, T. (2025). *Development of Automated Data Quality Assessment and Evaluation Indices by Analytical Experience* (No. arXiv:2504.02663). arXiv. https://doi.org/10.48550/arXiv.2504.02663
38. Holstein, J., Spitzer, P., Hoell, M., Vössing, M., & Kühl, N. (2024). *Understanding Data Understanding: A Framework to Navigate the Intricacies of Data Analytics*. Proceedings of the 32nd European Conference on Information Systems (ECIS 2024), Paphos, Cyprus.
39. Horani, O. M., Khatibi, A., AL-Soud, A. R., Tham, J., Al-Adwan, A. S., & Azam, S. M. F.

(2023). ANTECEDENTS OF BUSINESS ANALYTICS ADOPTION AND IMPACTS ON BANKS' PERFORMANCE: THE PERSPECTIVE OF THE TOE FRAMEWORK AND RESOURCE-BASED VIEW. *Interdisciplinary Journal of Information, Knowledge, and Management*, *18*, 609–643. Scopus. https://doi.org/10.28945/5188

40. *Informatica Data Quality and Observability*. (n.d.). Informatica. Retrieved May 1, 2025, from https://www.informatica.com/products/data-quality.html
41. Isaja, M., Nguyen, P., Goknil, A., Sen, S., Husom, E. J., Tverdal, S., Anand, A., Jiang, Y., Pedersen, K. J., Myrseth, P., Stang, J., Niavis, H., Pfeifhofer, S., & Lamplmair, P. (2023). A blockchain-based framework for trusted quality data sharing towards zero-defect manufacturing. *Computers in Industry*, *146*, 103853. https://doi.org/10.1016/j.compind.2023.103853
42. *ISO/IEC 25012:2008*. (n.d.). ISO. Retrieved May 1, 2025, from https://www.iso.org/standard/35736.html
43. Jackson, A. (2024, May 8). *Top 10: AI Tools for Data Analysis*. https://aimagazine.com/top10/top-10-ai-tools-for-data-analysis
44. Jeon, K.-C., Han, G.-S., Han, C.-Y., & Chong, I. (2023). Federated Learning Model for Contextual Sensitive Data Quality Applications: Healthcare Use Case. *2023 31st Signal Processing and Communications Applications Conference (SIU)*, 1–4. https://doi.org/10.1109/SIU59756.2023.10223768
45. Jiang, L., & Zhao, J. (2012). An empirical study on risk data quality management. *2012 International Conference on Information Management, Innovation Management and Industrial Engineering*, *1*, 511–514. https://doi.org/10.1109/ICIII.2012.6339714
46. *Julius AI | Your AI Data Analyst*. (n.d.). Retrieved May 2, 2025, from https://julius.ai
47. Kaplan, A., & Haenlein, M. (2019). Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons*, *62*(1), 15–25. https://doi.org/10.1016/j.bushor.2018.08.004
48. Kavak, M., & Rusu, L. (2025). *Challenges and opportunities of artificial intelligence in digital transformation: A systematic literature review*. *Procedia Computer Science, 256*, 369–377. https://doi.org/10.1016/j.procs.2025.02.132
49. Kitchenham, B., & Brereton, P. (2013). A systematic review of systematic review process research in software engineering. *Information and Software Technology*, *55*(12), 2049–2075. https://doi.org/10.1016/j.infsof.2013.07.010
50. Loshin, D. (2002). Rule-based data quality. *Proceedings of the Eleventh International Conference on Information and Knowledge Management*, 614–616. https://doi.org/10.1145/584792.584894
51. *Machine Learning Courses | Online Courses for All Levels | DataCamp*. (n.d.). Retrieved May 2, 2025, from https://www.datacamp.com/category/machine-learning?page=1
52. *Machine Learning Fundamentals in R | DataCamp*. (n.d.). Retrieved May 2, 2025, from https://www.datacamp.com/tracks/machine-learning-fundamentals
53. Malterud, K., Siersma, V. D., & Guassora, A. D. (2016). Sample size in qualitative interview studies: Guided by information power. Qualitative Health Research, 26(13), 1753–1760.
54. Naeem, M., Irfan, M., Ahmad, M. I., & Anwar, F. (2023). A step-by-step process of thematic analysis to develop a conceptual model in qualitative research. *Frontiers in Psychology, 14*, 1100730. https://doi.org/10.3389/fpsyg.2023.1100730
55. Nargesian, F., Zhu, E., Miller, R. J., Pu, K. Q., & Arocena, P. C. (2019). Data lake management: Challenges and opportunities. *Proceedings of the VLDB Endowment*, *12*(12), 1986–1989. https://doi.org/10.14778/3352063.3352116
56. Nascimento, D. C., Pires, C. E., & Mestre, D. (2016). Data quality monitoring of cloud databases based on data quality SLAs. In *Big-Data Analytics and Cloud Computing: Theory, Algorithms and Applications* (pp. 3–20). Scopus. https://doi.org/10.1007/978-3-319-25313-8_1
57. Nascimento, D. C., Pires, C. E., & Mestre, D. G. (2015). A data quality-aware cloud service based on metaheuristic and machine learning provisioning algorithms. *Proceedings of the 30th Annual ACM Symposium on Applied Computing*, 1696–1703. https://doi.org/10.1145/2695664.2695753

58. Necba, H., Rhanoui, M., & El Asri, B. (2018). Using unsupervised machine learning for data quality. Application to financial governmental data integration. *Communications in Computer and Information Science*, *872*, 197–209. Scopus. https://doi.org/10.1007/978-3-319-96292-4_16
59. Nikiforova, A. (2020). Definition and evaluation of data quality: User-oriented data object-driven approach to data quality assessment. Baltic Journal of Modern Computing, 8(3). https://doi.org/10.22364/bjmc.2020.8.3.02
60. Palinkas, L. A., Horwitz, S. M., Green, C. A., Wisdom, J. P., Duan, N., & Hoagwood, K. (2015). Purposeful sampling for qualitative data collection and analysis in mixed method implementation research. *Administration and Policy in Mental Health*, *42*(5), 533–544. https://doi.org/10.1007/s10488-013-0528-y
61. Payette, M., Abdul-Nour, G., Meango, T. J.-M., & Côté, A. (2023). Improving Maintenance Data Quality: Application of Natural Language Processing to Asset Management. *Lecture Notes in Mechanical Engineering*, 582–589. Scopus. https://doi.org/10.1007/978-3-031-25448-2_54
62. Pipino, L., Lee, Y., & Wang, R. (2003). Data Quality Assessment. *Communications of the ACM*, *45*. https://doi.org/10.1145/505248.506010
63. Price, M., Schroeder, P., Bautzer, T., Schroeder, P., & Bautzer, T. (2024, July 11). US regulators fine Citi $136 million for failing to fix longstanding data issues. *Reuters*. https://www.reuters.com/business/finance/us-bank-regulators-fine-citi-136-million-failing-address-longstanding-data-2024-07-10/
64. Purwanto, A., Zuiderwijk, A., & Janssen, M. (2020). *Citizens' trust in open government data*. 310–318. Scopus. https://doi.org/10.1145/3396956.3396958
65. *Quality Monitoring & Data Profiling Tool Development ☑ Acropolium's Case Study*. (n.d.). Retrieved May 1, 2025, from https://acropolium.com/portfolio/ai-powered-quality-monitoring-data-profiling-tool/
66. Raca, V., Velinov, G., Dzalev, S., & Kon-Popovska, M. (2022). A Framework for Evaluation and Improvement of Open Government Data Quality: Application to the Western Balkans National Open Data Portals. *Sage Open*, *12*(2), 21582440221104813. https://doi.org/10.1177/21582440221104813
67. Ranjit, S., & Kawaljeet, S. (2010). A Descriptive Classification of Causes of Data Quality Problems in Data Warehousing. *International Journal of Computer Science Issues*, *7*.
68. Rasool, T., & Warraich, N. F. (2018). *Does quality matter: A systematic review of information quality of E-government websites*. 433–442. Scopus. https://doi.org/10.1145/3209415.3209473
69. Redman, T. C. (2001). *Data Quality: The Field Guide*. Digital Press.
70. Rutledge, P. (2020). *In-Depth Interviews*. 1–7. https://doi.org/10.1002/9781119011071.iemp0019
71. *SAS: Data and AI Solutions*. (n.d.). Retrieved May 1, 2025, from https://www.sas.com/en_us/home.html
72. *School of data analysis*. (n.d.). School of Data Analysis. Retrieved May 1, 2025, from https://dataschool.yandex.com/
73. Schwabe, D., Becker, K., Seyferth, M., Klaß, A., & Schaeffter, T. (2024). The METRIC-framework for assessing data quality for trustworthy AI in medicine: A systematic review. *Npj Digital Medicine*, *7*(1), 1–30. https://doi.org/10.1038/s41746-024-01196-4
74. Serra, M., Fernández-Medina, E., Rosado, D. G., & Villagrá, V. A. (2024). *Use of context in data quality management: A systematic literature review*. Journal of Systems and Software, 206, 111717. https://doi.org/10.1016/j.jss.2023.111717
75. Shah, K. N., Gami, S. J., & Trehan, A. (2024). *An intelligent approach to data quality management: AI-powered quality monitoring in analytics*. International Journal of Advanced Research in Science, Communication and Technology, 4(3). https://doi.org/10.48175/IJARSCT-22820
76. Shahzad, T., Mazhar, T., Tariq, M. U., Ahmad, W., Ouahada, K., & Hamam, H. (2025). A comprehensive review of large language models: Issues and solutions in learning environments. *Discover Sustainability*, *6*(1), 27. https://doi.org/10.1007/s43621-025-00815-8

77. Siddiqa, A., Karim, A., & Gani, A. (2017). Big data storage technologies: A survey. *Frontiers of Information Technology & Electronic Engineering*, *18*(8), 1040–1070. https://doi.org/10.1631/FITEE.1500441
78. *Software*. (n.d.). Retrieved May 2, 2025, from https://www.hpe.com/asia_pac/en/software.html
79. Song, S., Gao, F., Huang, R., & Wang, C. (2023). Data Dependencies Extended for Variety and Veracity: A Family Tree (Extended abstract). *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, 3819–3820. https://doi.org/10.1109/ICDE55515.2023.00336
80. Soni, S., Marx, E., Katsavounidis, E., Essick, R., Cabourn Davies, G. S., Brockill, P., Coughlin, M. W., Ghosh, S., & Godwin, P. (2024). QoQ: A Q-transform based test for gravitational wave transient events. *Classical and Quantum Gravity*, *41*(1), 015012. https://doi.org/10.1088/1361-6382/ad0922
81. Tamm, H. C., & Nikiforova, A. (2025). *Towards AI-Augmented Data Quality Management: From Data Quality for AI to AI for Data Quality Management* (No. arXiv:2406.10940). arXiv. https://doi.org/10.48550/arXiv.2406.10940
82. Ustunboyacioglu, I., Kumara, I., Di Nucci, D., Tamburri, D. A., & Van Den Heuvel, W.-J. (2024). Data Quality Assessment in the Wild: Findings from GitHub. *Proceedings of the 28th International Conference on Evaluation and Assessment in Software Engineering*, 120–129. https://doi.org/10.1145/3661167.3661213
83. Ustunboyacioglu, I., Kumara, I., Di Nucci, D., Tamburri, D. A., & van den Heuvel, W.-J. (2024). Integrating Data Quality in Industrial Big Data Architectures: An Action Design Research Study. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *14889 LNCS*, 3–19. Scopus. https://doi.org/10.1007/978-3-031-70797-1_1
84. Véliz, C., Gkatzelis, V., & Loukis, E. (2024). Privacy and data protection challenges in generative AI systems: Risks and regulatory perspectives. *Computers & Security, 130*, 103295. https://doi.org/10.1016/j.cose.2023.103295
85. Vilella, J., Ratajczak, P., Sintsova, V., Lacasa, L., & Rovira-Asenjo, N. (2024). *Anomaly detection in cross-country money transfer temporal networks*. arXiv preprint arXiv:2311.14778. https://doi.org/10.48550/arXiv.2311.14778
86. Wang, R. Y. (1998). *A Product Perspective on Total Data Quality Management*. *Communications of the ACM, 41*(2), 58–65.
87. Wang, R. Y., & Strong, D. M. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems*, *12*(4), 5–33.
88. Wang, X., Li, X., & Xia, X. (2024). Research on Data Quality Management Methods and Technologies. *2024 IEEE 2nd International Conference on Image Processing and Computer Applications (ICIPCA)*, 116–120. https://doi.org/10.1109/ICIPCA61593.2024.10709151
89. Wang, Y., Song, S., Chen, L., Yu, J. X., & Cheng, H. (2017). Discovering Conditional Matching Rules. *ACM Transactions on Knowledge Discovery from Data*, *11*(4), 1–38. https://doi.org/10.1145/3070647
90. *Welcome—YData Profiling*. (n.d.). Retrieved May 1, 2025, from https://docs.profiling.ydata.ai/latest/
91. Whang, S. E., Roh, Y., Song, H., & Lee, J.-G. (2023). Data collection and quality challenges in deep learning: A data-centric AI perspective. *The VLDB Journal*, *32*(4), 791–813. https://doi.org/10.1007/s00778-022-00775-9
92. Zednik, C. (2021). Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence. *Philosophy & Technology*, *34*(2), 265–288. https://doi.org/10.1007/s13347-019-00382-7
93. Zhang, L., Howard, S., Montpool, T., Moore, J., Mahajan, K., & Miranskyy, A. (2023). Automated data validation: An industrial experience report. *Journal of Systems and Software*, *197*, 111573. https://doi.org/10.1016/j.jss.2022.111573
94. Zhang, Y., & Zhang, J. (2025). A comprehensive survey on generative AI: Architectures, training data, and interpretability challenges. *Artificial Intelligence Review*, *58*(1), 1–25. https://doi.org/10.1016/j.air.2025.01.004
95. Zhou, Y., Tu, F., Sha, K., Ding, J., & Chen, H. (2024). A Survey on Data Quality Dimensions

and Tools for Machine Learning Invited Paper. *2024 IEEE International Conference on Artificial Intelligence Testing (AITest)*, 120–131. https://doi.org/10.1109/AITest62860.2024.00023

96. Zong, X., & Vlachos, D. G. (2023). Reconciling experimental catalytic data stemming from structure sensitivity. *Chemical Science*, *14*(16), 4337–4345. Scopus. https://doi.org/10.1039/d2sc06819b

Appendices

Appendix I: Interview Protocol

Opening Statement

Thank you for participating in this study on the role of artificial intelligence (AI) in data quality management (DQM). This study is being conducted by Kaisa Käosaar under the supervision of Anastasija Nikiforova, both from the University of Tartu.

The purpose of this interview is to explore the challenges faced by data professionals in managing data quality and to understand the role that AI-based tools and solutions play in addressing these challenges. We aim to identify patterns, best practices and gaps in current data quality management approaches to inform future improvements and support the selection of AI-based data quality tools based on specific DQ challenges.

As with any research activity, there is a minimal risk of a breach of confidentiality. To the best of our ability, your answers in this study will remain confidential. We will minimize any risks by anonymizing the collected data, and only general characteristics (such as industry type) may be referenced in the analysis. All personal data, including your name and contact details, will be deleted within six months after data collection.

Your participation in this study is entirely voluntary, and you are free to withdraw at any time without any consequences. You may also choose to skip any questions you do not wish to answer. After the interview, you will have the opportunity to review the interview summary and request corrections or clarifications if necessary.

The interview will take approximately 1 hour to complete. We will be asking you 28 questions, beginning with general questions about your professional background and experience with data quality management. We will then proceed to more specific questions about the challenges you face in managing data quality, the tools and strategies you use, and the potential value and limitations of AI-based solutions in this area.

If you have any questions or concerns about this study, please feel free to contact:

Kaisa Käosaar - kaisa.kaosaar@gmail.com
Anastasija Nikiforova - nikiforova.anastasija@gmail.com

PLEASE TICK THE APPROPRIATE BOXES	Yes	No
A: GENERAL AGREEMENT – RESEARCH GOALS, PARTICIPANT TASKS AND VOLUNTARY PARTICIPATION		
1. I have read and understood the study information dated [*DD/MM/YYYY*], or it has been read to me. I have been able to ask questions about the study and my questions have been answered to my satisfaction.	☐	☐
2. I consent voluntarily to be a participant in this study and understand that I can refuse to answer questions and I can withdraw from the study at any time, without having to give a reason.	☐	☐
3. I understand that taking part in the study involves an audio-recorded interview, which is further transcribed, and once this is done, the audio recording will be destroyed.	☐	☐
4. I understand that the study will end after a sufficient number of interviews across different countries have been conducted to refine the model with its further application. The expected end date of this interview study is May/June 2025.	☐	☐
B: POTENTIAL RISKS OF PARTICIPATING (INCLUDING DATA PROTECTION)		
5. I understand that taking part in the study also involves collecting specific personally identifiable information (PII) (i.e., your name and signature on this consent form and your e-mail address if you decide to provide this to receive our study's results) and associated personally identifiable research data (PIRD) (i.e., your organization) with the potential risk of my identity being revealed.	☐	☐
6. I understand that the following steps will be taken to minimize the threat of a data breach, and protect my identity in the event of such a breach by anonymizing the data and mostly using them in an aggregated manner. Audio records will be transcribed (only accessible to the study team), while audio recordings will be deleted after transcription. Transcriptions will be anonymized and stored locally and secured. We will only report about organizations where interviewees are employed in an aggregated way and quotes in our publications will be anonymized.	☐	☐
7. I understand that personal information collected about me that can identify me, such as my name will not be shared beyond the study team.	☐	☐
8. I understand that the (identifiable) personal data I provide will be destroyed by the end of the research.	☐	☐
C: RESEARCH PUBLICATION, DISSEMINATION AND APPLICATION		

PLEASE TICK THE APPROPRIATE BOXES	Yes	No
9. I understand that after the research study the de-identified information I provide will be used for reports, publications, websites, as well as the development of policies to improve the current state of the art of government data sharing. In specific cases we expect to cite provided answers (anonymized quotes) to ensure accurate references to particular issues.	☐	☐
10. I agree that my responses, views or other input can be quoted anonymously in research outputs.	☐	☐

Signatures

_____ _____ _____

Name of participant [printed] Signature Date

I, as researcher, have accurately read out the information sheet to the potential participant and, to the best of my ability, ensured that the participant understands to what they are freely consenting.

_____ _____ _____

Researcher name [printed] Signature Date

Interview Questions

General

This section aims to understand your professional background and experience with data quality management (DQM). We want to get a sense of your expertise, how long you've been working in this field, and the types of tasks and stakeholders you typically work with.

Q1. What is your area of expertise?
Q2. How long have you been working with data quality management?
Q3. What are the DQM tasks you are typically involved in and for whom (e.g. businesses, governments, internal teams, other stakeholders)?
Q4. Data quality of which data domain(s)' do you primarily manage (e.g., finance, healthcare, manufacturing, domain-agnostic)?

Challenges

In this section, we will explore the key challenges you face in data quality management. We are interested in understanding both the general challenges you encounter and more specific difficulties related to different phases of DQM, data types, storage structures, and data quality dimensions. Your insights will help us identify patterns and gaps in current DQM practices.

Q5. What are the biggest challenges you currently encounter in managing data quality?
Q5.1. Have these challenges evolved over the last few years?
Q6. From your experience, what are the most common data quality challenges faced by non-DQ experts (e.g., business users, analysts)?
Q7. Which phases of DQM (definition, measurement, analysis, improvement) do you find most difficult to manage? What makes it difficult (e.g. is it time-consuming, human resource-intensive, requires business domain knowledge)?
Q8. Are there challenges related to data type *(e.g., structured, unstructured, big data, real-time data)*?
Q8.1. If yes, what specific issues arise when working with different types of data?
Q9. Are there challenges related to how your data is stored or structured *(e.g., databases, data lakes, data warehouses, data mesh)*?
Q9.1. If yes, what specific issues arise when working with this type of data storage/structure?
Q10. Are certain data quality dimensions more difficult to manage than others *(e.g., timeliness, accuracy, completeness, consistency etc.)*?
Q10.1. If yes, what specific issues arise when managing these dimensions?
Q11. Are there any underlying factors we haven't covered that make DQ management difficult?

Current Approach

This section focuses on how you currently manage data quality. We would like to know whether you rely on manual, semi-automated, or automated methods and which tools you use. We are also interested in understanding whether business domain knowledge is required to resolve issues and how communication and ownership affect the process.

Q12. How do you currently approach data quality?
- Manually
- Semi-automatically
- Automatically
Q13. What tools (if any) do you use today for DQ?
Q14. What tools (if any) have you used in the past for DQ?
Q15. In resolving data quality issues:
Q15.1. Is it enough to rely on DQ expertise alone, or is business domain knowledge (e.g., understanding business rules) required?
Q15.2. How often do you need to consult business experts or data owners to understand or resolve data quality issues?
Q15.3. Are there clear, predefined business rules in place for resolving these issues, or does it require case-by-case judgment?
Q15.4. Is communication between business and DQ teams a challenge when resolving these issues?

AI in Data Quality Management

In this section, we would like to explore your experience with AI-based data quality tools and processes. We are interested in learning whether you have applied AI to manage data quality, how effective it has been, and where you see the most value in using AI for DQM. We will also ask about potential barriers to adopting AI-based solutions.

Q16. Do you currently use AI in your data quality management processes or tools? ☐ If yes, go to question 17 ☐ If no, go to question 18
Q17. In what ways is AI being used in your data quality management processes?
Q18. What are the reasons for not using AI in your data quality management processes?
Q19. Where do you see AI providing the most value in DQM beyond how you're using it today?
Q20. Have you applied AI to address any of the specific challenges discussed earlier? ☐ If yes, go to question 21 ☐ If no, go to question 22
Q21. How effective has AI been in addressing those challenges?
Q22. What are the reasons for not using AI to address those challenges?
Q23. Do you think AI-based DQ tools should be designed primarily for DQ experts or business users?
Q23.1 Could AI reduce the need for technical expertise in DQ management?
Q24. What are the biggest issues with adopting AI-based DQ solutions:
- Technical complexity

- Lack of transparency - *You can see the outcome of the AI, but you don't know how the AI reached that conclusion*
- Lack of explainability - *Even if you know how the AI reached the conclusion, the logic or reasoning behind it isn't clear*
- Cost
- Trust in AI decision-making
- Integration with existing systems
- Regulatory or compliance issues
- Other (please specify)

Future

This section looks toward the future of data quality management. We would like to hear your thoughts on the potential impact of generative AI (GenAI) and other emerging technologies on DQM practices. We are particularly interested in whether you believe GenAI could improve data quality earlier in the data lifecycle and whether AI could play a broader role in data governance.

Q25. Do you see DQM to be a subject for change with an advancement of technologies, such as GenAI - does it hold potential to transform or at least somehow change the current DQM practices or have an effect on DQ expert daily routines?
Q26. Could GenAI improve the quality of data at earlier stages of the data lifecycle (before it reaches DQ experts)?
Q27. Do you think GenAI will significantly change how organizations handle DQM?
Q28. Do you see potential for AI to improve data governance beyond DQM?

Closing question	Yes	No
Would you like to get informed about the results of this study?	☐	☐

Thank you for your participation in this study! We appreciate your time!

Appendix II: Tool Evaluation Protocol

Section	Field	Description	Explanation
General Information	Tool Name	Name of the tool	Name of the tool being evaluated.
General Information	Type	Open-source / Commercial	Is the tool open-source, free with trial or commercial only?
General Information	Website	Link to official website	Link to the tool's official website.
General Information	Documentation Link	If available	Link to setup guides, API docs, etc.
General Information	Trial Available	Yes / No / (Notes)	Is there a free trial? If so, how long or limited?
General Information	Demo Video / Tour	Yes / No / (Notes)	Is there a video overview or product tour available?
AI Usage	AI Features Present	Yes / No / (Notes)	Does the tool use AI? Based on documentation, vendor claims or observed features (AI explicitly stated).
AI Usage	AI Type / Technique	ML / DL / Rule-based / NLP / Generative / Hybrid	What type of AI is used (Machine Learning, NLP, rules, etc.)? Based on documentation, observed functionality or vendor claims.
AI Usage	AI Location	Configurable / User-visible	Is AI used behind the scenes or configurable by the user?
AI Usage	Explainability	Black-box / White-box / Visuals / Logs	Are AI decisions accompanied by elements that support explainability (e.g., visuals, logs, rule tracing)?
AI Usage	AI Customizability	Pre-trained only / Trainable / Rule-extendable	Can users modify or extend the AI models or logic?
Functional Capabilities	Data Profiling	Yes / No / Partially / (Notes)	Can the tool analyse and summarize datasets (e.g., types, ranges, nulls)?
Functional Capabilities	DQ Rules (custom definition)	Yes / No / Partially / (Notes)	Can the user define their own data quality rules?
Functional Capabilities	SQL-based Rule Definition	Yes / No / Partially / (Notes)	Are rules definable using SQL?
Functional Capabilities	Predefined DQ Dimensions	Yes / No / Partially / (Notes)	Does the tool support traditional DQ dimensions (e.g., completeness, accuracy) in rule definition, profiling, or monitoring?
Functional Capabilities	Rules Repository	Yes / No / Partially / (Notes)	Is there a way to save the defined rules for their further reuse?
Functional Capabilities	Error Reporting	Yes / No / Partially / (Notes)	Can the tool identify and report invalid or incorrect data based on defined rules or fully automatically?
Functional Capabilities	DQ Dashboard	Yes / No / Partially / (Notes)	Can the tool provide a dashboard to monitor data quality over time?
Functional Capabilities	Match Detection	Yes / No / Partially / (Notes)	Can it detect duplicate or potentially matching records?
Functional Capabilities	Anomaly Detection	Yes / No / Partially / (Notes)	Can the tool automatically detect outliers or anomalies?
Functional Capabilities	Rule-based DQ Checks	Yes / No / Partially / (Notes)	Can the tool systematically validate data using built-in and/or user-defined rules?
Functional Capabilities	Data Cleansing	Yes / No / Partially / (Notes)	Can the tool perform data cleaning or error correction?
Functional Capabilities	Data Enrichment	Yes / No / Partially / (Notes)	Can the tool automatically fill in missing values or does it require manual input?
Functional Capabilities	Master Data Management	Yes / No / Partially / (Notes)	Does it support managing golden records / master data entities?
Functional	Data Lineage	Yes / No / Partially / (Notes)	Can it track where data came from and how it changed?

Section	Field	Description	Explanation
Capabilities			
Functional Capabilities	Data Catalogue	Yes / No / Partially / (Notes)	Is there a searchable catalogue of datasets and metadata?
Functional Capabilities	Semantic Discovery	Yes / No / Partially / (Notes)	Can it automatically identify meanings or relationships in data?
Functional Capabilities	Data Integration	Yes / No / Partially / (Notes)	Does it support connecting and combining data from multiple sources?
Usability & Workflow Fit	User Interface	Modern / Legacy / CLI-based	What kind of UI does it have? CLI-based, legacy, or modern graphical UI?
Usability & Workflow Fit	Designed for Business Users	Yes / No / Partially	Can non-technical users operate the tool without coding or technical setup?
Usability & Workflow Fit	Ease of Onboarding	Moderate / Good	How extensive are the supportive materials for the tool (Docs, tutorials, support)?
Usability & Workflow Fit	Workflow Integration	API-first / GUI flows / No-code pipelines	Can it fit into existing pipelines (APIs, drag-and-drop, etc.)?
Usability & Workflow Fit	Real-Time / Batch Support	Real-time / Batch only	Does it process data in real time, batches or both?
Usability & Workflow Fit	Data Sources Supported	Moderate / Good	Types of systems it can connect to (e.g. SQL, APIs, etc.).
Usability & Workflow Fit	Cloud Compatibility	Cloud-native / On-prem only	Can it run on or integrate with cloud platforms (e.g., AWS / Azure / GCP)?
Usability & Workflow Fit	Collaboration Features	Comments / Shared rules / Team permissions	Are there team features like sharing, permissions, comments?
Data Processing	JSON	Supported / Partially Supported / Not supported	Can it read/write JSON files (used in APIs and web data)?
Data Processing	CSV, TSV	Supported / Partially Supported / Not supported	Can it handle flat files like CSV/TSV (spreadsheet-like formats)?
Data Processing	Relational Databases	Supported / Partially Supported / Not supported	Can it connect to SQL databases like MySQL, Postgres, Oracle?
Data Processing	NoSQL Databases	Supported / Partially Supported / Not supported	Does it support non-SQL databases like MongoDB, Cassandra?
Data Processing	Spreadsheets	Supported / Partially Supported / Not supported	Can it handle Excel files or Google Sheets?
Data Processing	Data Lakes	Supported / Partially Supported / Not supported	Does it work with big data storage like S3 or HDFS?
Data Processing	Data Warehouses	Supported / Partially Supported / Not supported	Can it connect to systems like Snowflake, BigQuery etc?
Data Processing	APIs	Supported / Partially Supported / Not supported	Can it pull or push data via APIs?
Data Processing	Cloud-native Support	Supported / Partially Supported / Not supported	Is the tool designed to be deployed and operated in cloud environments (per documentation or platform support)?

Appendix III: Full List of Tools Identified through Google Search

Source type	Source Title	Tools Mentioned in Source	Reference
Article	AI Magazine. "Top 10: AI Tools for Data Analysis"	1. RapidMiner 2. Tableau 3. Qlik 4. Polymer 5. Databricks Unified Data Analytics Platform 6. Sisense 7. The KNIME Analytics Platform 8. IBM Watson Analytics 9. Google Cloud Smart Analytics 10. Microsoft Azure Machine Learning	(Jackson, 2024)
Homepage	SAS Homepage	SAS	(*SAS*, n.d.)
Homepage	HPE Homepage	HPE	(*Software*, n.d.)
Homepage	Julius AI Homepage	Julius AI	(*Julius AI | Your AI Data Analyst*, n.d.)
Homepage	Informatica Homepage	Informatica	(*Informatica Data Quality and Observability*, n.d.)
Python Package	YData Profiling Documentation	YData Profiling	(*Welcome - YData Profiling*, n.d.)
Homepage	Ataccama Homepage	Ataccama	(*Effortless Data Quality, Infused with AI*, n.d.)
Research Paper	The METRIC-framework for assessing data quality for trustworthy AI in medicine: a systematic review	-	(Schwabe et al., 2024)
Homepage	Acropolium Homepage	Acropolium	(*Quality Monitoring & Data Profiling Tool Development ☑ Acropolium's Case Study*, n.d.)
Homepage	DIGNA AI Homepage	DIGNA AI	(*Digna - AI-Powered Data Quality for Data Warehouses & Co., Made in Europe*, n.d.)
Article	"Top 12 AI Tools for Data Analysis To Include In Your Tech Stack"	1. RapidMiner 2. Talend 3. ThoughtSpot	(AnalytixLabs, 2025)

Source type	Source Title	Tools Mentioned in Source	Reference
		4. KNIME 5. Google Sheets 6. DataRobot 7. Akkio 8. IBM Watson Analytics 9. H20.ai 10. Microsoft Power BI 11. Tableau 12. Luzmo	
Course	DataCamp - Machine Learning Fundamentals in R	-	(*Machine Learning Fundamentals in R | DataCamp*, n.d.)
Course	Yandex School of Data Analysis	-	(*School of Data Analysis*, n.d.)
Course	DataCamp - Machine Learning Courses	-	(*Machine Learning Courses | Online Courses for All Levels | DataCamp*, n.d.)

Appendix IV: Full List of Tools Identified through Scopus Search

Article name	Author(s)	Year	Source	Tools Mentioned in Source	Reference
A design theory for data quality tools in data ecosystems: Findings from three industry cases	Altendeitering M.; Guggenberger T.M.; Möller F.	2024	Data and Knowledge Engineering	-	(Altendeitering et al., 2024)
Data Quality Assessment in the Wild: Findings from GitHub	Ustunboyacioglu I.; Kumara I.; Di Nucci D.; Tamburri D.A.; Van Den Heuvel W.-J.	2024	ACM International Conference Proceeding Series	1. (Py)Deequ 2. Pandera 3. Data Build Tool (DBT) 4. Great Expectations (GX) 5. TensorFlow Data Validation (TFDV)	(Ustunboyacioglu, Kumara, Di Nucci, Tamburri, & Van Den Heuvel, 2024)
QoQ: a Q-transform based test for gravitational wave transient events	Soni S.; Marx E.; Katsavounidis E.; Essick R.; Cabourn Davies G.S.; Brockill P.; Coughlin M.W.; Ghosh S.; Godwin P.	2024	Classical and Quantum Gravity	-	(Soni et al., 2024)
A Survey on Data Quality Dimensions and Tools for Machine Learning Invited Paper	Zhou Y.; Tu F.; Sha K.; Ding J.; Chen H.	2024	Proceedings - 6th IEEE International Conference on Artificial Intelligence Testing, AITest 2024	1. Kylo 2. MobyDQ 3. Apache Griffin 4. SQL Power Architect 5. Aggregate Profiler 6. YData Quality 7. DataCleaner 8. WinPure 9. SQL Power DQguru 10. Deequ 11. Dataedo 12. OpenRefine 13. Great Expectations 14. Soda 15. Ataccama ONE 16. whylogs 17. Evidently	(Zhou et al., 2024)
Integrating Data Quality in Industrial Big Data Architectures: An Action Design Research Study	Ustunboyacioglu I.; Kumara I.; Di Nucci D.; Tamburri D.A.; van den Heuvel W.-J.	2024	Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)	1. PyDeequ 2. AWS	(Ustunboyacioglu, Kumara, Di Nucci, Tamburri, & van den Heuvel, 2024)
Evaluation and Optimization of Machine Learning Algorithms in Personalized Marketing	Guo C.; Jiao P.	2024	3rd IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics, ICDCECE 2024	-	(Guo & Jiao, 2024)

Article name	Author(s)	Year	Source	Tools Mentioned in Source	Reference
Data Quality Tools: Towards a Software Reference Architecture	Altendeitering M.; Guggenberger T.M.	2024	Proceedings of the Annual Hawaii International Conference on System Sciences	-	(Altendeitering & Guggenberger, 2024)
Context-aware Big Data Quality Assessment: A Scoping Review	Fadlallah H.; Kilany R.; Dhayne H.; El Haddad R.; Haque R.; Taher Y.; Jaber A.	2023	Journal of Data and Information Quality	1. QuaIIe 2. SparkDQ	(Fadlallah et al., 2023)
A blockchain-based framework for trusted quality data sharing towards zero-defect manufacturing	Isaja M.; Nguyen P.; Goknil A.; Sen S.; Husom E.J.; Tverdal S.; Anand A.; Jiang Y.; Pedersen K.J.; Myrseth P.; Stang J.; Niavis H.; Pfeifhofer S.; Lamplmair P.	2023	Computers in Industry	-	(Isaja et al., 2023)
Reconciling experimental catalytic data stemming from structure sensitivity	Zong X.; Vlachos D.G.	2023	Chemical Science	-	(Zong & Vlachos, 2023)
Factors Influencing Practitioners' Experience of Utilizing Open Government Data	Cheng W.-C.; Chiu M.-H.P.	2023	Journal of Library and Information Studies	-	(Cheng & Chiu, 2023)
Federated Learning Model for Contextual Sensitive Data Quality Applications: Healthcare Use Case	Jeon K.-C.; Han G.-S.; Han C.-Y.; Chong I.	2023	31st IEEE Conference on Signal Processing and Communications Applications, SIU 2023	-	(Jeon et al., 2023)
ANTECEDENTS OF BUSINESS ANALYTICS ADOPTION AND IMPACTS ON BANKS' PERFORMANCE: THE PERSPECTIVE OF THE TOE FRAMEWORK AND RESOURCE-BASED VIEW	Horani O.M.; Khatibi A.; AL-Soud A.R.; Tham J.; Al-Adwan A.S.; Azam S.M.F.	2023	Interdisciplinary Journal of Information, Knowledge, and Management	-	(Horani et al., 2023)
Automating Data Quality Monitoring with Reference Data Profiles	Ehrlinger L.; Werth B.; Wöß W.	2023	Communications in Computer and Information Science	1. DQ-MeeRKat 2. HoloDetect	(Ehrlinger et al., 2023)
Improving Maintenance Data Quality: Application of Natural Language Processing to Asset Management	Payette M.; Abdul-Nour G.; Meango T.J.-M.; Côté A.	2023	Lecture Notes in Mechanical Engineering	-	(Payette et al., 2023)
Data Dependencies Extended for Variety and Veracity: A Family Tree	Song S.; Gao F.; Huang R.; Wang C.	2022	IEEE Transactions on Knowledge and Data Engineering	-	(Song et al., 2023)
A Framework for Evaluation and Improvement of Open Government Data Quality: Application to the Western Balkans National Open Data Portals	Raca V.; Velinov G.; Dzalev S.; Kon-Popovska M.	2022	SAGE Open	-	(Raca et al., 2022)
Data Quality in Data Ecosystems: Towards a Design Theory	Altendeitering M.; Dübler S.; Guggenberger T.	2022	28th Americas Conference on Information Systems, AMCIS 2022	-	(Altendeitering, Dübler, et al., 2022)

Article name	Author(s)	Year	Source	Tools Mentioned in Source	Reference
Data Sovereignty for AI Pipelines: Lessons Learned from an Industrial Project at Mondragon Corporation	Altendeitering M.; Pampus J.; Larrinaga F.; Legaristi J.; Howar F.	2022	Proceedings - 1st International Conference on AI Engineering - Software Engineering for AI, CAIN 2022	-	(Altendeitering, Pampus, et al., 2022)
Improving knowledge based detection of soft attacks against autonomous vehicles with reputation, trust and data quality service models	Chuprov S.; Viksnin I.; Kim I.; Melnikov T.; Reznik L.; Khokhlov I.	2021	Proceedings - 2021 IEEE International Conference on Smart Data Services, SMDS 2021	-	(Chuprov et al., 2021)
Data science training for official statistics: A new scientific paradigm of information and knowledge development in national statistical systems	Ashofteh A.; Bravo J.M.	2021	Statistical Journal of the IAOS	-	(Ashofteh & Bravo, 2021)
DQ-MeeRKat: Automating data quality monitoring with a reference-data-profile-annotated knowledge graph	Ehrlinger L.; Gindlhumer A.; Huber L.-M.; Wöß W.	2021	Proceedings of the 10th International Conference on Data Science, Technology and Applications, DATA 2021	-	(Ehrlinger et al., 2021)
Citizens' trust in open government data	Purwanto A.; Zuiderwijk A.; Janssen M.	2020	ACM International Conference Proceeding Series	-	(Purwanto et al., 2020)
How to inspect and measure data quality about scientific publications: Use case of Wikipedia and CRIS databases	Azeroual O.; Lewoniewski W.	2020	Algorithms	DataCleaner	(Azeroual & Lewoniewski, 2020)
Does quality matter: A systematic review of information quality of E-government websites	Rasool T.; Warraich N.F.	2018	ACM International Conference Proceeding Series	-	(Rasool & Warraich, 2018)
Using unsupervised machine learning for data quality. Application to financial governmental data integration	Necba H.; Rhanoui M.; El Asri B.	2018	Communications in Computer and Information Science	-	(Necba et al., 2018)
Discovering conditional matching rules	Wang Y.; Song S.; Chen L.; Yu J.X.; Cheng H.	2017	ACM Transactions on Knowledge Discovery from Data	-	(Y. Wang et al., 2017)
Challenges for value-driven semantic data quality management	Brennan R.	2017	ICEIS 2017 - Proceedings of the 19th International Conference on Enterprise Information Systems	1. Dacura Quality Service 2. RDFUnit	(Brennan, 2017)
Describing reasoning results with RVO, the reasoning violations ontology	Božić B.; Brennan R.; Feeney K.C.; Mendel-Gleason G.	2016	CEUR Workshop Proceedings	Dacura Quality Service	(Božić et al., 2016)
Data quality monitoring of cloud databases based on data quality SLAs	Nascimento D.C.; Pires C.E.; Mestre D.	2016	Big-Data Analytics and Cloud Computing: Theory, Algorithms and Applications	-	(Nascimento et al., 2016)

Article name	Author(s)	Year	Source	Tools Mentioned in Source	Reference
A data quality-aware cloud service based on metaheuristic and machine learning provisioning algorithms	Nascimento D.C.; Pires C.E.; Mestre D.G.	2015	Proceedings of the ACM Symposium on Applied Computing	DQaS	(Nascimento et al., 2015)
Data quality in big data: A review	Abdullah N.; Ismail S.A.; Sophiayati S.; Sam S.M.	2015	International Journal of Advances in Soft Computing and its Applications	-	(Abdullah et al., n.d.)

Appendix V: Tool Evaluation Protocol (General Information and AI Usage)

Tool Name	Type	Website	Documentation Link	Trial Available	Demo Video /Tour	AI Features Present	AI Type / Technique	AI Location	Explainability	AI Customizability
RapidMiner	Commercial	https://altair.com/altair-rapidminer	https://support.altair.com/csm?id=altair_product_documentation	Yes	Yes	Yes	Hybrid (Machine Learning, Deep Learning, NLP)	Configurable / User-visible	White-box / Visuals (for some models)	Trainable / Rule-extendable
Tableau	Commercial	https://www.tableau.com	https://www.tableau.com/support	Yes	Yes	Yes	Hybrid (Generative AI / NLP / ML)	Configurable / User-visible	Visuals	Pre-trained only
Qlik	Commercial	https://www.qlik.com/us	https://help.qlik.com/en-US/cloud-services/Content/Sense_Helpsites/Home.htm	Yes	Yes	Yes	Hybrid (ML / NLP / Generative)	User-visible	Visuals	Pre-trained / Extendable (via AutoML)
Polymer	Commercial	https://www.polymersearch.com	https://apidocs.polymersearch.com/	Yes	Yes	Yes	NLP / ML	User-visible	Visuals	Pre-trained only

Tool Name	Type	Website	Documentation Link	Trial Available	Demo Video /Tour	AI Features Present	AI Type / Technique	AI Location	Explainability	AI Customizability
DataBricks	Commercial	https://www.databricks.com	https://docs.databricks.com/aws/en	Yes	Yes	Yes	Hybrid (Machine Learning, Deep Learning, NLP)	Configurable / User-visible	White-box / Visuals / Logs (for some models)	Trainable / Extendable
Sisense	Commercial	https://www.sisense.com	https://docs.sisense.com/	Yes	Yes	Yes	Hybrid (Machine learning, NLP)	Configurable / User-visible	Visuals	Limited Direct Customization
KNIME	Open-source (offers commercial extensions)	https://www.knime.com	https://docs.knime.com	N-A – Core platform is free; commercial server edition available on request	Yes	Yes	Hybrid (Machine Learning, Deep Learning, NLP)	User-visible / Configurable	White-box / Visuals / Logs (for some models)	Trainable / Rule-extendable
IBM Watson Analytics	Commercial	https://www.ibm.com/products/watson-studio	https://www.ibm.com/docs/en?lnk=flathl	Yes	Yes	Yes	ML / NLP / Generative	Configurable / User-visible	White-box / Visuals	Trainable / Rule-extendable
Google Cloud Smart Analytics	Comemercial	https://cloud.google.com	https://cloud.google.com/docs	Yes	Yes	Yes	Hybrid (ML / DL /	Configurable /	Black-box / White-box / Visuals / Logs	Pre-trained / Trainable

Tool Name	Type	Website	Documentation Link	Trial Available	Demo Video /Tour	AI Features Present	AI Type / Technique	AI Location	Explainability	AI Customizability
							Generative / NLP)	User-visible		
Microsoft Azure Machine Learning	Commercial	https://azure.microsoft.com/	https://learn.microsoft.com/en-us/azure/?product=popular	Yes	Yes	Yes	Hybrid	Configurable / User-visible	White-box / Visuals / Logs	Trainable / Rule-extendable
SAS Vija	Commercial	https://www.sas.com/	https://support.sas.com/en/documentation.html	Yes	Yes	Yes	Hybrid (ML / NLP / Generative)	Configurable / User-visible	White-box / Visuals / Logs	Trainable / Rule-extendable
HPE Machine Learning Data Management	Commercial	https://www.hpe.com/au/en/software.html	https://docs.ai-solutions.ext.hpe.com/products/mldm/	No	Yes	No	-	-	-	-
Julius AI	Commercial	https://julius.ai	https://julius.ai/docs/get-started/welcome	Yes	Yes	Yes	Hybrid (ML / NLP / Generative)	User-visible	Visuals	Pre-trained only

Tool Name	Type	Website	Documentation Link	Trial Available	Demo Video /Tour	AI Features Present	AI Type / Technique	AI Location	Explainability	AI Customizability
Informatica	Commercial	https://www.informatica.com	https://docs.informatica.com	Yes	Yes	Yes	Hybrid (Machine Learning, NLP, Rules)	Configurable and User-visible	Visuals / Logs	Trainable / Rule-Extendable
YData Profiling	Open-source	https://docs.profiling.ydata.ai/latest/	https://docs.profiling.ydata.ai/latest/getting-started/concepts/	N-A - free to use	No	No	-	-	-	-
Ataccama	Commercial	https://www.ataccama.com	https://docs.ataccama.com/	No	Yes	Yes	Hybrid (Machine Learning, Generative AI, Rule-based)	Configurable and User-visible	Visuals, Logs, and Natural Language Explanations	Trainable and Rule-extendable
Acropolium	Commercial	https://acropolium.com/	-	No	No	-	-	-	-	-
Digna AI	Commercial	https://www.digna.ai	-	No	No	-	-	-	-	-
Talend	Commercial	www.talend.com	https://help.talend.com	Yes	Yes	Yes	ML / Generative	User-visible and Configurable	Visuals / Logs	Trainable

Tool Name	Type	Website	Documentation Link	Trial Available	Demo Video /Tour	AI Features Present	AI Type / Technique	AI Location	Explainability	AI Customizability
ThoughtSpot	Commercial	https://www.thoughtspot.com	https://docs.thoughtspot.com/home	Yes	Yes	Yes	Hybrid (DL / NLP / ML)	User-visible	Visuals	Pre-trained
Google Sheets	Free	https://sheets.google.com/	https://support.google.com/docs	N-A - free to use	Yes	No	-	-	-	-
DataRobot	Commercial	https://www.datarobot.com	https://docs.datarobot.com/en/docs/index.html	Yes	Yes	Yes	Hybrid (ML / NLP / Generative)	Configurable / User-visible	White-box / Visuals / Logs	Pre-trained / Trainable / Rule-extendable
Akkio	Commercial	https://www.akkio.com	https://docs.akkio.com/akkio-docs	Yes	Yes	Yes	Hybrid (ML / NLP / Generative)	User-visible	Visuals	Pre-trained
H2O	Open-source / Commercial	https://h2o.ai	https://docs.h2o.ai/	Yes	Yes	Yes	Hybrid (ML / DL / NLP / Generative)	Configurable / User-visible	White-box / Visuals / Logs	Trainable and Partially Rule-Extendable

Tool Name	Type	Website	Documentation Link	Trial Available	Demo Video /Tour	AI Features Present	AI Type / Technique	AI Location	Explainability	AI Customizability
Microsoft Power BI	Commercial	https://powerbi.microsoft.com	https://learn.microsoft.com/en-gb/power-bi/	Yes	Yes	Yes	Hybrid (ML / NLP)	Configurable (limited) / User-visible	Black-box / White-box / Visuals / Logs	Trainable
Luzmo	Commercial	https://www.luzmo.com	https://developer.luzmo.com	Yes	Yes	Yes	NLP / Generative	User-visible	Visuals	Pre-trained only
Kylo	Open-source	https://kylo.io	https://kylo.readthedocs.io/en/v0.10.0/	N-A - free to use	Yes	-	-	-	-	-
MobyDQ	Open-source	https://ubisoft.github.io/mobydq/	-	N-A - free to use	No	No	-	-	-	-
Apache Griffin	Open-source	https://griffin.apache.org	https://griffin.apache.org/docs/quickstart.html	N-A - free to use	No	No	-	-	-	-
SQL Power Architect	Open-source	https://bestofbi.com/products/sql-power-architect-data-modeling/	No	N-A - free to use	Yes	No	-	-	-	-
Aggregate Profiler	Open-source	https://sourceforge.net/projects/dataquality/	-	N-A - free to use	No	No	-	-	-	-

Tool Name	Type	Website	Documentation Link	Trial Available	Demo Video /Tour	AI Features Present	AI Type / Technique	AI Location	Explainability	AI Customizability
DataCleaner	Open-source	https://datacleaner.github.io	https://datacleaner.github.io/docs/5.7.0/html/	N-A - free to use	No	No	-	-	-	-
WinPure	Commercial	https://winpure.com	https://docs.winpure.com/	Yes	Yes	Yes	ML / Rule-based	Not directly user-configurable	Visuals	Rule-extendable
SQLPowerDQGuru	Open-source	https://bestofbi.com/products/sql-power-dqguru-data-quality/	No	N-A - free to use	Yes	No	-	-	-	-
Dataedo	Commercial	https://dataedo.com	https://docs.dataedo.com	Yes	Yes	Yes	Hybrid (Generative / NLP)	Configurable / User-visible	White-box / Visuals	Pre-trained
Openrefine	Open-source	https://openrefine.org	https://openrefine.org/docs	N-A – free to use	No	No	-	-	-	-
Soda	Hybrid (Soda Core is open-source, Soda Cloud is commercial)	https://soda.io	https://docs.soda.io	Yes	Yes	Yes	Generative / NLP	User-visible /configurable	Visual	No (can only prompt the model)

Tool Name	Type	Website	Documentation Link	Trial Available	Demo Video /Tour	AI Features Present	AI Type / Technique	AI Location	Explainability	AI Customizability
whylogs	Hybrid – Open-source (library) / Commercial Cloud (WhyLabs)	https://whylabs.ai	https://whylogs.readthedocs.io/	Yes	Yes	No	-	-	-	-
Evidently AI	Hybrid – Open-source / Commercial Cloud	https://www.evidentlyai.com	https://docs.evidentlyai.com/	Yes	Upon request	Yes	ML / LLM	Configurable / User-visible	Visuals	Rule-extendable
(Py)Deequ	Open-source	https://github.com/awslabs/deequ	https://github.com/awslabs/deequ/blob/master/README.md *read-me	N-A - free to use	No	No	-	-	-	-
Pandera	Open-source	https://www.union.ai/pandera	https://pandera.readthedocs.io/en/stable/#	N-A - free to use	No	No	-	-	-	-
Data Build Tool (dbt)	Hybrid – Open-source (DBT Core) + Commercial (DBT Cloud)	https://www.getdbt.com	https://docs.getdbt.com	Yes	Yes	Yes	Hybrid (NLP / Generative)	Configurable / User-visible	White-box / Visuals	Trainable

Tool Name	Type	Website	Documentation Link	Trial Available	Demo Video /Tour	AI Features Present	AI Type / Technique	AI Location	Explainability	AI Customizability
Great Expectations (GX)	Open-source	https://greatexpectations.io	https://docs.greatexpectations.io	N-A - free to use	Yes	No	-	-	-	-
Tensorflow Data Validation (TFDV)	Open-source	https://www.tensorflow.org	https://www.tensorflow.org/guide#essential-documentation	N-A - free to use	Yes	Yes	Hybrid (ML / DL / NLP / Generative)	Configurable / User-visible	White-box / Visuals / Logs	Trainable / Rule-extendable
QuaIIe	Open-source	https://github.com/zbw/qualle	-	N-A - free to use	No	No	-	-	-	-
SparkDQ	Open-source	https://github.com/PasaLab/SparkDQ	-	N-A - free to use	No	No	-	-	-	-
DQ-MeeRKat	Open-source	https://github.com/lisehr/dq-meerkat	https://github.com/lisehr/dq-meerkat/tree/master/documentation	N-A - free to use	No	No	-	-	-	-
HoloDetect	Academic / Research Prototype	https://arxiv.org/abs/1904.02285	Not available	No	No	Yes	ML	Configurable / User-visible	Logs	Trainable
Dacura Quality Service	-	-	-	No	No	No	-	-	-	-

Tool Name	Type	Website	Documentation Link	Trial Available	Demo Video /Tour	AI Features Present	AI Type / Technique	AI Location	Explainability	AI Customizability
RDFUnit	Open-source	https://github.com/AKSW/RDFUnit	https://github.com/AKSW/RDFUnit/wiki	N-A - free to use	No	No	-	-	-	-
Oracle Enterprise Data Quality (EDQ)	Commercial	https://www.oracle.com/	https://docs.oracle.com/	Yes	Yes	No	-	-	-	-

Appendix VI: Tool Evaluation Protocol (Functionalities)

Tool Name	Data Profiling	DQ Rules (custom definition)	SQL-based Rule Definition	Predefined DQ Dimensions	Rules Repository	Error Reporting	DQ Dashboard	Match Detection	Anomaly Detection	Rule-based DQ Checks	Data Cleansing	Data Enrichment	Master Data Management	Data Lineage	Data Catalogue	Semantic Discovery	Data Integration
RapidMiner	Yes	Yes	Yes	No	No	Yes	Partially (through reporting extensions)	Yes	Yes	Partially	Yes	Yes	No	Partially	No	No	Yes
Tableau	Yes	Yes	Yes	No	No	Yes	Yes	No	Yes	Yes	Yes	Partially (through joins)	No	Yes	Yes	Partially	Yes
Qlik	Yes	Yes	Yes	No	Partially (No formal repository)	Yes	Yes	No	Yes	Yes	Yes	Partially (through joins and lookups)	No	Partially (Within a Qlik application)	Yes (add-on)	Yes	Yes
DataBricks	Yes	Yes	Yes	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Partially	Yes	Yes	No	Yes
Sisense	Yes	Yes	Partially	No	No	Yes	Yes	No	Yes	Yes	Yes	Yes	No	Partially	No	Partially (through Sisense NLQ)	Yes

Tool Name	Data Profiling	DQ Rules (custom definition)	SQL-based Rule Definition	Predefined DQ Dimensions	Rules Repository	Error Reporting	DQ Dashboard	Match Detection	Anomaly Detection	Rule-based DQ Checks	Data Cleansing	Data Enrichment	Master Data Management	Data Lineage	Data Catalogue	Semantic Discovery	Data Integration
KNIME	Yes	Yes	Yes	Partially	Partially	Yes	Partially	Yes	Yes	Yes	Yes	Yes	No	Yes	No	Not directly, can be done hrough AI/ML Integrations	Yes
IBM Watson Analytics	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Google Cloud Smart Analytics	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Partially	Yes	Yes	Partially	Yes
Microsoft Azure Machine Learning	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
SAS Vija	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Julius AI	Yes	No	No	No	No	No	No	No	Yes	No	Yes	No	No	No	No	No	Yes

Tool Name	Data Profiling	DQ Rules (custom definition)	SQL-based Rule Definition	Predefined DQ Dimensions	Rules Repository	Error Reporting	DQ Dashboard	Match Detection	Anomaly Detection	Rule-based DQ Checks	Data Cleansing	Data Enrichment	Master Data Management	Data Lineage	Data Catalogue	Semantic Discovery	Data Integration
Informatica	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Ataccama	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Talend	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
ThoughtSpot	Yes	No	No	No	No	Partially	Partially	No	Yes	No	No	No	No	No	Yes	Yes	Yes
DataRobot	Yes	No	No	Yes	No	Yes	Yes	No	Yes	No	Yes	No	No	Partially	Yes	Partially	Yes
Akkio	Yes	No	No	No	No	Partially	No	No	Yes	No	Yes	No	No	No	No	Yes	Yes
H2O	Yes	Partially (Through custom recipes and constraints)	No	No	Partially	Yes	Yes	Yes	Yes	Partially	Yes	Yes	No	No	No	Yes	Yes

Tool Name	Data Profiling	DQ Rules (custom definition)	SQL-based Rule Definition	Predefined DQ Dimensions	Rules Repository	Error Reporting	DQ Dashboard	Match Detection	Anomaly Detection	Rule-based DQ Checks	Data Cleansing	Data Enrichment	Master Data Management	Data Lineage	Data Catalogue	Semantic Discovery	Data Integration
Microsoft Power BI	Yes	Yes	Yes	Partially	No	Yes	Yes	No	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes
Luzmo	Yes	No	No	No	No	Partially	Partially	No	Yes	No	Partially	No	No	No	No	Partially	Yes
WinPure	Yes	Yes	No	Partially	Yes	Yes	Partially	Yes	Yes	Yes	Yes	Yes	No	No	No	Yes	Yes
Dataedo	Yes	Yes	Yes	Partially	Yes	Yes	Yes	No	No	Yes	No	No	Partially	Yes	Yes	Partially	Yes
Soda	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	No	No	No	No	No (but supports integration with external catalogs)	No	No
Evidently AI	Yes	Yes	No	No	No	Yes	Yes	No	Partially	Yes	No	No	No	No	No	No	Partially

Tool Name	Data Profiling	DQ Rules (custom definition)	SQL-based Rule Definition	Predefined DQ Dimensions	Rules Repository	Error Reporting	DQ Dashboard	Match Detection	Anomaly Detection	Rule-based DQ Checks	Data Cleansing	Data Enrichment	Master Data Management	Data Lineage	Data Catalogue	Semantic Discovery	Data Integration
Data Build Tool (dbt)	Partially	Yes	Yes	Partially	Yes	Yes	Partially	No	No	Yes	Yes	Yes	No	Yes	Yes	Partially	No
Tensorflow Data Validation (TFDV)	Partially (via extensions)	Partially (requires implementa)	No	No	No	Yes	No	Not directly, but can be implemented	Yes	Partially	Partially	Yes	No	No	No	Partially	Yes
HoloDetect	No	No	No	No	No	No	No	No	No	No	Yes	Yes	No	No	No	No	No
Polymer	Yes	No	No	No	No	No	Yes	No	Yes	No	No	No	No	No	No	No	Yes

Appendix VII: Tool Evaluation Protocol (Usability and Data Processing)

Tool Name	User Interface	Designed for Business Users	Ease of Onboarding	Workflow Integration	Real-Time / Batch Support	Data Sources Supported	Cloud Compatibility	Collaboration Features	JSON	CSV, TSV	Relational Databases	NoSQL Databases	Spreadsheets	Data Lakes	Data Warehouses	APIs	Cloud-native Support
RapidMiner	Modern	Partially	Documentation / Tutorials / Support quality - Good	GUI flows / API (via extensions and scripting)	Batch / Real-time (through extensions)	e.g. SQL, NoSQL, Flat files, APIs - Extensive	AWS / Azure / GCP / On-prem	Comments / Shared rules / Team permissions	Supported	Supported	Supported	Supported (via extensions)	Supported	Supoorted (via connectors)	Supported (via connectors)	Supported	Supported
Tableau	Modern	Yes	Documentation / Tutorials / Support quality - Good	API-first / GUI flows	Near Real-time / Batch	e.g. SQL, NoSQL, Flat files, APIs - Extensive	AWS / Azure / GCP / On-prem	Comments / Shared rules / Team permissions	Supported	Supported	Supported	Supported	Supported	Supported	Supported	Supported	Supported
Qlik	Modern	Yes	Documentation / Tutorials / Support quality - Good	API-first / GUI flows	Near Real-time / Batch	e.g. SQL, NoSQL, Flat files, APIs - Extensive	AWS / Azure / GCP / On-prem	Comments / Shared rules / Team permissions	Supported	Supported	Supported	Supported	Supported	Supported	Supported	Supported	Supported
DataBricks	Modern (Web-based)	Partially	Documentation / Tutorials / Support quality - Good	API-first / GUI flows	Streaming / Batch	e.g. SQL, NoSQL, Flat files, APIs - Extensive	Cloud-native (AWS / Azure / GCP)	Comments / Shared rules / Team permissions	Supported	Supported	Supported	Supported	Supported	Supported	Supported	Supported	Supported
Sisense	Modern	Partially	Documentation / Tutorials / Support quality - Good	API-first / GUI flows	Batch only	e.g. SQL, NoSQL, Flat files, APIs - Extensive	AWS / Azure / GCP / On-prem	Comments / Shared rules / Team permissions	Supported	Supported	Supported	Supported	Supported	Supported	Supported	Supported	Supported

Tool Name	User Interface	Designed for Business Users	Ease of Onboarding	Workflow Integration	Real-Time / Batch Support	Data Sources Supported	Cloud Compatibility	Collaboration Features	JSON	CSV, TSV	Relational Databases	NoSQL Databases	Spreadsheets	Data Lakes	Data Warehouses	APIs	Cloud-native Support
KNIME	Modern	Partially	Documentation / Tutorials / Support quality - Good	GUI flows / API (via extensions)	Batch primarily / Real-time through extensions	e.g. SQL, NoSQL, Flat files, APIs - Extensive	AWS / Azure / GCP / On-prem	Comments / Shared rules / Team permissions - Partially available (via KNIME Server/Business Hub)	Supported	Supported	Supported	Supported (via extensions)	Supported	Supported (via connectors)	Supported	Supported	Supported
IBM Watson Analytics	Modern	Partially	Documentation / Tutorials / Support quality - Good	API-first / GUI flows	Real-time / Batch	e.g. SQL, NoSQL, Flat files, APIs - Extensive	Cloud-native / On-prem	Shared rules / Team permissions	Supported	Supported	Supported	Supported	Supported	Supported	Supported	Supported	Supported
Google Cloud Smart Analytics	Modern	Partially	Documentation / Tutorials / Support quality - good	API-first / GUI flows / No-code pipelines	Real-time / Batch	e.g. SQL, NoSQL, Flat files, APIs - Extensive	Cloud-native / On-prem	Shared rules / Team permissions	Supported	Supported	Supported	Supported	Supported	Supported	Supported	Supported	Supported
Microsoft Azure Machine Learning	Modern / CLI-based	Partially	Documentation / Tutorials / Support quality - Good	API-first / GUI flows / No-code pipelines	Real-time / Batch	e.g. SQL, NoSQL, Flat files, APIs - Extensive	Cloud-native / On-prem	Shared rules / Team permissions	Supported	Supported	Supported	Supported	Supported	Supported	Supported	Supported	Supported
SAS Vija	Modern	Partially	Documentation / Tutorials /	API-first / GUI flows	Real-time / Batch	e.g. SQL, NoSQL, Flat files, APIs	AWS / Azure / GCP / On-prem only	Comments / Shared rules /	Supported	Supported	Supported	Supported	Supported	Supported	Supported	Supported	Supported

Tool Name	User Interface	Designed for Business Users	Ease of Onboarding	Workflow Integration	Real-Time / Batch Support	Data Sources Supported	Cloud Compatibility	Collaboration Features	JSON	CSV, TSV	Relational Databases	NoSQL Databases	Spreadsheets	Data Lakes	Data Warehouses	APIs	Cloud-native Support
			Support quality - Good					Team permissions									
Julius AI	Modern	Yes	Documentation / Tutorials / Support quality - Good	API-first /GUI Flows/ No-code pipelines	Batch	Limited (Flat files, Google Sheets, Postgres)	Cloud compatible (no mention of specific providers)	Team permissions	Supported	Supported	Supported	Not supported	Supported	Not supported	Not supported	Not supported	Partially supported
Informatica	Modern	Yes	Documentation / Tutorials / Support quality - Good	API-first / GUI flows / No-code pipelines	Real-time / Streaming / Batch only	e.g. SQL, NoSQL, Flat files, APIs - Extensive	AWS / Azure / GCP / On-prem	Comments / Shared rules / Team permissions	Supported	Supported	Supported	Supported	Supported	Supported	Supported	Supported	Supported
Ataccama	Modern	Partially	Documentation / Tutorials / Support quality - Good	GUI flows / API-first	Real-time / Streaming / Batch	Source agnostic	AWS / Azure / GCP / On-prem	Comments / Shared rules / Team permissions	Supported	Supported	Supported	Supported	Supported	Supported	Supported	Supported	Supported
Talend	Modern	Yes	Documentation / Tutorials / Support quality - Good	API-first, GUI flows	Real-time / Batch	e.g. SQL, NoSQL, Flat files, APIs - Extensive	Cloud-nati / On-prem	Shared rules / Team permissions	Supported	Supported	Supported	Supported	Supported	Supported	Supported	Supported	Supported
ThoughtSpot	Modern	Yes	Documentation / Tutorials /	API-first	Real-time / Batch	e.g. SQL, NoSQL, Flat	Cloud-native / On-prem	Shared dashboards/	Supported	Supported	Supported	Supported (partially)	Supported	Supported	Supported	Supported	Supported

Tool Name	User Interface	Designed for Business Users	Ease of Onboarding	Workflow Integration	Real-Time / Batch Support	Data Sources Supported	Cloud Compatibility	Collaboration Features	JSON	CSV, TSV	Relational Databases	NoSQL Databases	Spreadsheets	Data Lakes	Data Warehouses	APIs	Cloud-native Support
			Support quality - Good			files, APIs - Extensive		Team permissions									
DataRobot	Modern	Yes	Documentation / Tutorials / Support quality - Good	API-first	Real-time / Batch	e.g. SQL, NoSQL, Flat files, APIs- Extensive	Cloud-native	Shared rules / Team permissions	Supported	Supported	Supported	Supported	Supported	Supported	Supported	Supported	Supported
Akkio	Modern	Yes	Documentation / Tutorials / Support quality - Good	API-first	Batch	e.g. SQL, NoSQL, Flat files, APIs - Extensive	Cloud-native	Team permissions	Supported	Supported	Supported	Supported	Supported	Not supprted	Supported	Supported	Supported
H2O	Modern	Partially	Documentation / Tutorials / Support quality - Good	API-first / GUI flows / No-code pipelines	Real-time / Batch	e.g. SQL, NoSQL, Flat files, APIs - Extensive	AWS / Azure / GCP / On-prem	Shared rules / Team permissions	Supported	Supported	Supported	Supported	Supported	Supported	Supported	Supported	Supported
Microsoft Power BI	Modern	Yes	Documentation / Tutorials / Support quality - Good	API-first / GUI flows	Real-time / Batch	e.g. SQL, NoSQL, Flat files, APIs - Extensive	Cloud-native / On-prem	Comments / Shared rules / Team permissions	Supported	Supported	Supported	Supported	Supported	Supported	Supported	Supported	Supported
Luzmo	Modern	Yes	Documentation / Tutorials / Support quality - Good	API-first	Batch only	e.g. SQL, NoSQL, Flat files, APIs - Extensive	Cloud-native	Shared dashboards	Supported	Supported	Supported	Supported	Supported	Supported	Supported	Supported	Supported

Tool Name	User Interface	Designed for Business Users	Ease of Onboarding	Workflow Integration	Real-Time / Batch Support	Data Sources Supported	Cloud Compatibility	Collaboration Features	JSON	CSV, TSV	Relational Databases	NoSQL Databases	Spreadsheets	Data Lakes	Data Warehouses	APIs	Cloud-native Support
WinPure	Modern	Yes	Documentation / Tutorials / Support quality - Limited	GUI flows	Batch only	e.g. SQL, NoSQL, Flat files, APIs - Moderate	On-prem only	Team permissions	Not supported	Supported	Supported	Not supported	Supported	Not supported	Supported	Supported	Not supported
Dataedo	Modern	Partially	Documentation / Tutorials / Support quality - Good	API-first / GUI flows	Batch only	e.g. SQL, NoSQL, Flat files, APIs - Extensive	Cloud-native / On-prem	Shared rules, Team permissions	Supported	Supported	Supported	Supported	Supported	Supported	Supported	Supported	Supported
Soda	Modern	Yes	Good - offers documentation, Slack support, tutorials etc	API-first	Batch only	SQL Databases	Yes - cloud-native	Yes	Supported	Supported	Supported	Not supported	Not supported	Supported	Supported	Not supported	Supported
Evidently AI	Code-centric (Python library) with a Modern Graphical UI (Cloud Platform)	Partially	Documentation / Tutorials / Support quality - Good	API-first / GUI flows / No-code pipelines	Real-time / Batch	e.g. SQL, NoSQL, Flat files, APIs - Moderate	AWS / Azure / GCP / On-prem	Team permissions (on the Cloud Platform)	Supported	Supported	Not supported	Not supported	Not supported	Not supported	Not supported	Not supported	Supported

Tool Name	User Interface	Designed for Business Users	Ease of Onboarding	Workflow Integration	Real-Time / Batch Support	Data Sources Supported	Cloud Compatibility	Collaboration Features	JSON	CSV, TSV	Relational Databases	NoSQL Databases	Spreadsheets	Data Lakes	Data Warehouses	APIs	Cloud-native Support
Data Build Tool (dbt)	Modern / CLI-based	No	Documentation / Tutorials / Support quality - Good	API-first / CLI flows	Batch only	e.g. SQL, NoSQL, Flat files, APIs - Limited (Warehouse-focused)	Cloud-native / On-prem	Team permissions	Partially supported (dependent on data warehouse JSON support)	Not supported	Supported	Not supported	Not supported	Partially supported	Supported	Not supported	Supported
Tensorflow Data Validation (TFDV)	Code-centric	No	Documentation / Tutorials / Support quality - Good	Code-based	Batch / Real-Time	Extensive (via Python libraries)	AWS / Azure / GCP / On-prem	Limited (via external platforms)	Supported	Supported	Supported (via external libraries)	Supported (via external libraries)	Supported	Supported (via external libraries)	Supported (via external libraries)	Supported	Supported
HoloDetect	CLI-based	No	Low	Not available	Batch	CSV	Not specified	No	Supported	Supported	Not supported	Not supported	Not supported	Not supported	Not supported	Not supported	Not supported
Polymer	Modern	Yes	Documentation / Tutorials / Support quality - Good	API-first / GUI flows	Real-time	e.g. SQL, NoSQL, Flat files, APIs - Extensive	Cloud-native	Shared dashboards	Supported	Supported	Supported	Not supported	Supported	Not supported	Not supported	Supported	Supported

Licence

Non-exclusive licence to reproduce the thesis and make the thesis public

I, Kaisa Käosaar ,
(author´s name)

grant the University of Tartu a free permit (non-exclusive licence) to

reproduce, for the purpose of preservation, including for adding to the digital archives of the University of Tartu until the expiry of the term of copyright, my thesis

Exploring Data Quality Management Challenges and the Emerging Role of AI Solutions ,
(title of thesis)

supervised by Anastasija Nikiforova ;
(supervisor´s name)

2. grant the University of Tartu a permit to make the thesis specified in point 1 available to the public via the web environment of the University of Tartu, including via the digital archives, under the Creative Commons licence CC BY NC ND 4.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright;

3. am aware of the fact that the author retains the rights specified in points 1 and 2;

4. confirm that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Kaisa Käosaar
15/05/2025