Data Quality in Distributed Information Systems: Conceptual Challenges and Empirical Findings

Sven Vujević
University of Zagreb Faculty of Organization and Informatics
Pavlinska 2, 42000 Varaždin, Croatia
svujevic21@foi.hr

Maja Cerjan, Kornelije Rabuzin
University of Zagreb Faculty of Organization and Informatics
Laboratory for Data Technologies
Pavlinska 2, 42000 Varaždin, Croatia
{macerjan,krabuzin}@foi.unizg.hr

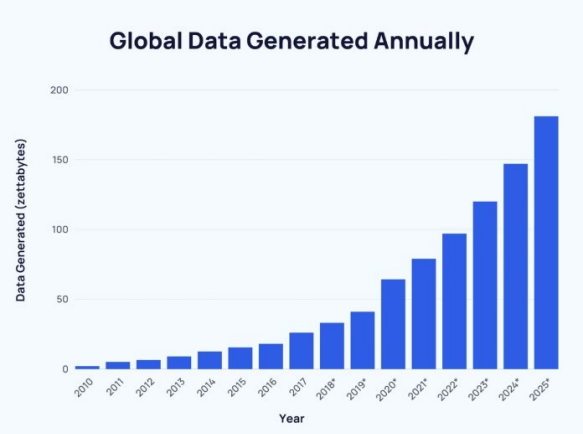
Abstract. *As is widely known by now, the quality of data is of great importance. However, conundrums arise if the data cleansing process isn't done appropriately from the very source. Especially in the digital world, driven by data in which its amounts are growing exponentially (Duarte, F., 2025). This paper emphasizes the importance of the quality of data in distributed information systems where the situation is even more perplexing and harder to resolve providing an error occurs.*
Therewithal, you will find the theoretical guidelines of things to do and ones to avoid, complying with ISO/IEC 25012 standard, either you are trying to get the grasp of your own data or you're working for an external business customer. To shed light on this growing problem there will also be concrete, firsthand examples from industry showing some of the most common fallacies and therefore challenges to overcome to ensure the quality of data for reliable decision making in the future.
This reading provides best practices when it comes to handling the data optimally both in theory and practice wherefore it is a great foundation for both field experts and field experts to be.

Keywords. data quality, distributed systems, data consistency, information systems, data cleansing, ISO/IEC 25012

1 Introduction

Regardless of the sector the company operates in, data-driven decisions have become the norm of business excellence and are inevitable if your goal is to stay competitive in the market. "Consequently, along with people, data can actually be considered as one of the most important assets for organizations" (Gualo, F. et al., 2021). It is trivial that data and its quality play an immense and crucial part in the process of decision making (Wang, R.Y., 1998), (Woodall, P. et al., 2012). According to IBM Data quality measures how well a dataset meets criteria for accuracy, completeness, validity, consistency, uniqueness, timeliness and fitness for purpose, and it is critical to all data governance initiatives within an organization. Furthermore, according to research by consulting firm Gartner, poor data quality costs companies an average of $12.9 million per year, while in the financial sector this amount can be significantly higher, and this is just one problematic example of this underlying problem. This only magnifies with the shire volume of ever new and exponentially bigger amounts of data being recorded every year (Duarte, F., 2025). "According to the latest estimates, 402.74 million terabytes of data are created each day" (Duarte, F., 2025).

Figure 1. Global data generated annually (Duarte, F., 2025)



This paradigm with the data as a cornerstone of the organizational activity and its success implies that the companies must fully commit to ensure the highest standards for the best outcomes, and the practice bolsters it (Gualo, F. et al., 2021). For example, when American Express began to foster more the importance of data quality with the accent on customer data accuracy, fraud detection, and behavioral analytics it had a huge positive impact which can be divided into three main parts: (1) delivering greater security for customers with 60% reduction in fraud by utilizing big data; (2) faster and smarter decisions backed with real-time data intelligence (3) long-term investment in innovation thus building a data-driven advantage (Llano, M., 2022). Another example of getting the best out of it is the story of Netflix. "We have embedded this approach into our culture from when the company was founded and call it Consumer (Data) Science. Broadly speaking, the main goal of Consumer Science

is to effectively innovate for users by using data to drive product decisions." (Amatriain, X., & Basilico, J., 2015). All this comes as motivation to invest your time and resources in studying this topic and implementing your findings in practice.

The sole process of ensuring high standards when it comes to data quality is composite and delicate work. There exist a lot of concepts and frameworks on this topic, some of which include the ISO/IEC 25012 standard that defines 15 dimensions that are nested across three different categories. By ISO/IEC 25012:2008 they are as follows: (1) Inherent; (2) Inherent and System-Dependent and (3) System-Dependent (Guerra-García, C. et al., 2023), (Miller, R. et al., 2024). Another approach is the DAQUAVORD methodology and its "data quality by design" that embeds data quality requirements into the early stages of information system development. DAQUAVORD is a relatively new concept (introduced in early 2023 by Guerra-García, C. et al.), nonetheless, based on the ISO/IEC 25012 standard.

In the following section one can find deeper overview of other relevant work done on this topic.

2 Related Work

2.1 Industrial Certification (ISO 25012)

Data quality has become a core asset for organizations, driving the need for systematic evaluation and certification (Gualo, F. et al., 2021). The ISO/IEC 25012 standard emerged as the gold standard and became central to data quality certification, offering a structured framework for evaluating and managing data quality within information systems (Gualo, F. et al., 2021). Initially, organizations recognized the importance of data quality for operational, tactical, and strategic activities (Gualo, F. et al., 2021). Poor data quality led to significant losses and hindered business initiatives, emphasizing the need for mechanisms to ensure data reliability (Gualo, F. et al., 2021). Just one of many such tremendous examples are Redman reported losses of about $3,1 billion for American companies due to poor and inadequate data quality, and the ever-growing complexity in system architecture which accounts for their high price (Redman, T. C., 2016), (Ballou, D. P., & Tayi, G. K., 1989).

Rather than treating all quality issues the same, international standards as ISO/IEC 25012 suggest evaluating data from multiple angles. Some aspects of quality stem from the data itself, such as whether it's correct or complete, while others depend on how systems manage, store, or present that data. By recognizing this distinction, organizations can take a more strategic approach by improving the data at its source while also ensuring that technological systems maintain their integrity throughout their lifecycle.

Apart from ISO/IEC 25012, other important standards include ISO/IEC 25024 and ISO/IEC 25040 standards, which are both part of the broader SQuaRE (Software Product Quality Requirements and Evaluation) series developed by ISO and IEC. They are helping to ensure quality in software and data systems in a broader, more holistic approach. ISO/IEC 25024:2015 defines a set of data quality measures for evaluating the quality of data used in systems and thus supporting the ISO/IEC 25012. Its purpose is to provide practical metrics and methods for measuring the actual quality of datasets, whether for internal auditing, improvement, certification, or regulatory purposes. Additionally, it defines quantitative metrics for each characteristic. It proves to be useful for data profiling and quality audits among others. ISO/IEC 25040:2024 defines a generic, structured process for evaluating software product quality, from planning to result reporting, based on requirements and context. It fulfils its purpose as repeatable evaluation framework that can be applied consistently across software products to determine whether they meet quality goals. It often works alongside ISO/IEC 25010:2023, which defines the quality model for software (e.g., reliability, usability, performance). Benefits of data quality certification include long-term organizational sustainability, better internal knowledge of data, and more efficient data quality management (Gualo, F. et al., 2021).

2.2 Alternative Certifications and Improvements on ISO 25012

However, challenges of data quality remain, including a lack of standardized terminology and the need for adaptable frameworks across different domains (Miller, R. et al., 2024). Recent work suggests enhancing the ISO 25012 standard with additional dimensions like governance, usefulness, quantity, and semantics to improve its applicability and relevance (Miller, R. et al., 2024). (1) Governance covers internal roles, authority, and accountability structures guiding data handling; (2) usefulness measures the extent to which data meets user/application needs, focusing on adaptability and reusability; (3) quantity considers whether there's a sufficient volume of data for a complete and accurate representation and (4) semantics captures the ability of data to convey correct and consistent meaning, crucial in contexts like ontologies or knowledge graphs (Miller, R. et al., 2024). This evolution aims to create a universal framework that supports effective data management and decision-making across various sectors (Miller, R. et al., 2024).

Another effort to improve upon ISO/IEC 25012 standard comes from Guerra-García et al. (2023) in their paper "ISO/IEC 25012-based methodology for managing data quality requirements in the development of information systems: Towards Data Quality by Design". Authors proposed structured methodology that significantly extends and operationalizes the ISO/IEC 25012:2008 standard. The solution included: (1) Integration of Data Quality into

Software Development Life Cycle (SDLC) - which encompassed the DAQUAVORD methodology, which embeds data quality requirements directly into the requirements engineering phase of system development (Guerra-García, C. et al., 2023). This shift supports the principle of "Data Quality by Design", ensuring that data quality concerns are proactively addressed, not retrofitted, which authors strongly encouraged (Guerra-García, C. et al., 2023). (2) Conceptual Extension is another such addition to current earlier discussed standard; it provides transformation from DQ dimensions to DQRs (Data Quality Requirements) and DQSRs (Data Quality Software Requirements) (Guerra-García, C. et al., 2023). Other key components are (3) Viewpoint-Oriented Requirements Definition (VORD) Integration; (4) Step-by-Step Methodology and Role Assignments (DAQUAVORD framework introduces a six-phase process) and (5) Realistic Use Case Implementation (the methodology is demonstrated using a university student registration system case study), (Guerra-García, C. et al., 2023).

2.3 Data Quality in Distributed Information Systems

Data quality in distributed data networks involves unique challenges compared to traditional, centralized systems due to the variety of data sources and the complexity of integrating data across multiple institutions (Kahn, M. G. et al. 2015). This implies additional steps in the process of data cleansing and establishing a sufficient level of the quality of data. Therefore, authors have been proposing different approaches to handling the challenge.

Some examples of the mentioned differences between the centralized and distributed systems include differences in a few key areas.

Table 1. Differences over data ownership and autonomy (Kahn, M. G. et al. 2015)

Centralized Systems:	**Distributed systems:**
Single point of control over data management	Multiple autonomous data providers/sites
Easier to enforce uniform data quality polices	Quality assessment varies per site and may lack transparency

Another significant difference between these two system architectures are the tools used to build, maintain and manipulate them. For example, for ETL process (Extract, Transform, Load) in centralized systems, the most common tools used are Informatica and Talend while for distributed systems some of the most used tools include Apache Spark, Kafka and Delta Lake among others (Margara, A. et al. 2022).

Table 2. Differences over cleansing and reconciliation complexity (Margara, A. et al. 2022)

Centralized Data Cleansing:	**Distributed Data Cleansing:**
Cleansing rules apply uniformly	Each node may apply its own logic or delay synchronization
Easier duplicate and inconsistency detection	Harder duplicate and inconsistency detection

Table 3. Quality assessment reporting (Kahn, M. G. et al. 2015)

Traditional systems:	**Distributed systems:**
Single-point metrics (accuracy, completeness)	Requires multi-site, temporal and source-specific indicators
Errors detected globally	Errors may be site-specific, temporal

3 Theoretical Framework: Dimensions of Data Quality

3.1 Data quality dimensions by ISO 25012 standard

There are multiple theoretical frameworks describing the dimensions of data quality in IT. However, some of those frameworks proved to be more credible than others over time. One such example is certainly the ISO/IEC 25012 standard.

ISO/IEC 25012:2008 standard describes data across 15 different dimensions which could be broadly arranged in three categories (Guerra-García, C. et al., 2023), (Miller, R. et al., 2024):

I. **Inherent**
II. **Inherent and System Dependent**
III. **System Dependent**

Inherent Data Quality Dimensions are the same regardless of the usage of the data (Guerra-García, C. et al., 2023), (Miller, R. et al., 2024). They assess the quality independent of the system in which the data resides, and the key idea is, if moved to another system, these qualities would remain the same, there are 5 of them and they are as follows: (1) accuracy, (2) completeness, (3) consistency, (4) credibility and (5) currentness.

Inherent and System-Dependent Data Quality Dimensions, which means they are hybrid of both inherent to a certain degree as well as partially dependent on the system they are used in (Guerra-García, C. et al., 2023), (Miller, R. et al., 2024). Their quality is determined by how well the data and the system work in cohesion: (6) accessibility, (7) compliance, (8) confidentiality, (9) efficiency, (10) precision, (11) traceability and (12) understandability.

System-Dependent Data Quality Dimensions, which means that they are different based on the nature and setting they are being used in (Guerra-García, C. et al., 2023), (Miller, R. et al., 2024). Given the different technical environments, they would always be the

same as before, these qualities change if the system changes even if the data stays the same: (13) availability, (14) portability and (15) recoverability.

Wang & Strong (1996) define quality dimensions as "a set of data quality attributes that represent a single aspect or construct of data quality". Data quality dimensions provide a means to quantify and manage data quality. When defining data quality measures, one should try to focus on dimensions that are meaningful and relevant to the business, with maximum return on investment. On the other hand, measuring all the different dimensions of data quality provides a complete picture. This is the motivation to go in more detail for the five core dimensions of data quality listed in the next subchapter.

3.2 Five core dimensions of data quality

The analysis of many studies showed that **completeness, accuracy, timeliness, consistency and relevance** are the first five dimensions of data quality that are most often mentioned in studies, so they will be described in more detail just below (Wang, R. Y. & Strong, D. M., 1996).

3.2.1 Completeness

Completeness is the most basic dimension in the family of data quality dimensions. Completeness is a measure of the presence or absence of data. In a relational database, "present data" means non-empty values in a data field in a table; "absent data" means null or empty values in a data field in a table. Sometimes values such as "unknown" or "not applicable" are also used to represent missing data (Mahanti, R., 2018). Typical questions that need to be asked are (Mahanti, R., 2018):

1. Is all the necessary information available?
2. Are critical values missing from the data records?
3. Are all data sets recorded?
4. Are all mandatory data recorded?

In some cases, missing data or information is irrelevant. However, in cases where missing data or information is critical to a particular business process or task being performed, completeness becomes a concern (Mahanti, R., 2018). Missing optional data is fine for data completeness. For example, a customer's middle name is optional, so the record can be considered complete even if the middle name is not available.

The percentage of missing/present data can be measured not only at the data item/element level but also at the record, dataset or database level. Missing records can have a huge impact, e.g. lost opportunities, additional costs, customer dissatisfaction, etc.

3.2.2 Accuracy

Accuracy generally means that the recorded value is consistent with a real-world fact or value. Accuracy refers to the absence of error and is considered by consumers to be the most important characteristic of data quality (Fisher, C. W., & Kingma, B. R., 2001). According to Wang & Strong (1996.) accuracy refers to the degree to which the data accurately represent the actual value or condition they describe. This includes the match between the data and the real world, that is, how similar the data is to the real entities they represent (Wang, R. Y. & Strong, D. M., 1996).

3.2.3 Timeliness

Timeliness refers to how up to date the data is for the task being performed. Data timeliness can be expressed as a function of: (1) how current the data is for the task for which it is being used, and (2) whether the data is available in time for use (Wang, R. Y. & Strong, D. M., 1996). According to Ballou and Pazer (1985), timeliness refers to whether information is available when it is needed. The value of timeliness decreases with age or as information changes due to new discoveries (Ballou, D. P., & Pazer, H. L., 1985). Data that does not comply with timeliness is often the result of processing delays or insufficiently frequent updates (Umar, A. et al. 1999). Outdated or delayed data can lead to inaccurate analyses or missed opportunities. For example, in the financial sector, timeliness is crucial when trading stocks. If an investor makes decisions based on stock price data that are not up to date, there is a risk of buying or selling at unfavorable conditions, which can result in financial losses.

3.2.4 Consistency

Consistency, by ISO 8000-8:2015 is a property of data that describes the degree to which the data is free of contradictions and consistent with other data in a specific context of use. Data consistency assesses the absence of obvious contradictions and discrepancies in the data set and the extent to which the data conforms to defined business rules, formats, and domains (Umar, A. et al. 1999). Consistency refers to the absence of contradictions in the data. For example, data that a customer is 20 years old but also has historical transactions recorded 25 years ago represents an inconsistency (Ballou, D. P., & Pazer, H. L., 1985).

Batini and Scannapieco (2006) expand on this concept: Consistency refers to the preservation and satisfaction of semantic rules defined over a set of data. Consistency can be defined in terms of models, constraints, and rules specific to application domains (Batini, C. & Scannapieco, M., 2006). Redman (2001) emphasizes the importance of consistency for business: Inconsistent data can lead to erroneous conclusions, duplicate efforts, and suboptimal decisions. Data consistency is a fundamental prerequisite for the integrity of an information system and the reliability of business processes (Redman, T. C., 2001).

3.2.5 Relevance

Relevance is a dimension of data quality that refers to the usefulness of the data for a specific task. It describes how well the information meets the needs of the user in a specific context of use (Kahn, B. K. et al., 2002). The ISO/IEC 25012:2008 standard defines relevance as the degree to which data has attributes that are appropriate and provide added value for a specific task in a specific context of use.

Relevance refers to the degree to which data is appropriate and useful for the specific task or decision at hand. It ensures that only the data that directly contributes to the objectives of a business process or analytical activity is collected, maintained, and used. High-quality data must not only be accurate and complete but also tailored to meet the needs of its intended users. Irrelevant data increases noise, storage costs, and processing time, and may even lead to flawed decision-making if it distracts from key indicators.

These five dimensions work synergistically: a deficiency in any one of them significantly reduces the overall quality of data. For example, complete but inaccurate data will lead to erroneous conclusions, accurate but outdated data may be useless for current decisions, timely but inconsistent data creates distrust, while consistent but irrelevant data wastes resources without creating value.

These dimensions also form the basis for most data quality assessment methodologies and standards such as ISO 8000 and DAMA-DMBOK, further confirming their central importance in data quality management.

4 Empirical Analysis of Data Quality Challenges

The foundation for the empirical analysis of the data quality challenges discussed in this academic paper is firsthand experience in the industry. Through this chapter there will be real observed examples of data quality issues, their causes and consequences and the way they were both discovered and handled in each specific scenario. We will try to present the wide array of different and unique challenges relating to the theoretical concepts described in the previous sections to connect the importance of understanding the theory and to be able to apply it accordingly.

The project was about data analysis for the purposes of fraud detection modelling for a big company within the SAS (Statistical Analysis System) environment. The analysis was carried out through SQL queries over DWH (data warehouse) (Oracle database) where two layers were covered: (1) The Dimensional Layer – in which the data model used for internal reporting was already built and (2) The Stage Layer – in which the original tables were mapped 1 to 1 due to the need for additional data that is not in the dimensional model.

4.1 Example 1: Inconsistent labelling (referencing) of the same entity (document) (violated dimension: consistency)

In one source system, a document is uniquely defined by a combination of four columns, while in the second system a unique ID is used, without the availability of all four columns that make up the key in the first system. This difference prevents simple linking of records between systems. It was detected by analysis through SQL queries on tables from both systems. Due to the impossibility of linking, it was decided to use data from only one system, which has more complete and significant information for the project.

Apart from the violation of the **consistency** (differing representations of the same entity across systems), this problem also directly affected other important dimensions. This is most usually the case since rarely the specific problem violates only one dimension because of their close interconnection. Other violations from the five core dimensions include violating: **completeness** (the second system lacks the four columns used in the composite key in the first system) and **relevance** (if the second system lacks detailed or contextual, then some of it may not be sufficiently relevant to the analytical needs). There are also multiple other dimensions violated from the ISO/IEC 25012:2008 standard as well.

4.2 Example 2: Different practice of recording data on entities (violated dimension: relevance)

We have 2 entities: (1) individuals who have acquired a certain right and (2) individuals who are using that right (users). Some data, although for the user of the right, are entered under the ID of the person who has acquired the right, and some data are entered under the ID of the person who is the user, so for each data item/table, it is necessary to separately check with the system users under which ID the data is entered. It was detected through the analysis of data obtained by SQL queries and those displayed in the application where there was a discrepancy, and through communication with the system users, an inconsistency was observed. Through communication with the system users, it was defined for which data item/table which ID is viewed.

In this example, apart from the violation of the **relevance** dimension (the data is not reliably associated with the correct person, reducing its usefulness in this context), **consistency** (the same type of data is recorded under different entity IDs across tables or cases) and **accuracy** (data may be technically correct in content (e.g., a date or status), but linked to the wrong person) were also violated as well as multiple dimensions from ISO/IEC 25012:2008 data quality standard.

4.3 Example 3: Differences between application view and DWH model – data inconsistency (violated dimension: accuracy)

The entitlement start date differs between the application and the DWH model, as DWH does not include an additional table of changes. The date calculation logic is based on a combination of multiple sources that the DWH model does not fully cover. The differences were observed through the analysis of the obtained data and the presentation in the application part. Due to the limitations of the DWH dimensional model (which is not under our control), it was decided to abandon the use of the DWH model for that part and use the stage part, which is a copy of the original table from the operating systems, and a clear rule for determining the mentioned date was defined with the end users.

Violated dimensions include **accuracy** (the DWH contains an incorrect or misleading start date), **completeness** (the DWH is missing necessary data (the change-tracking table)) and **consistency** (the same business concept appears with different values in the application and the DWH) as well as multiple dimensions from ISO/IEC 25012:2008 data quality standard.

4.4 Example 4: Missing data within the DWH system (violated dimension: completeness)

Continuing the previous problem where some dates are calculated from multiple entered dates, in addition to the insufficiently detailed published rule, there is also the problem that not all dates that should be considered exist within the DWH system. One such example was for the service start date where information on the service resignation date was missing. Without this information, it can be mistakenly considered that the service has started. The differences were observed through the analysis of the obtained data and the display in the application part. The analysis with the system users determined that the scope of the missing data is relatively small and does not significantly affect the results, so it was decided to continue the implementation without this data, with a documented assumption.

Violated dimensions include **completeness** (not all necessary data is present in the DWH (e.g., deletion records), which are critical for correct interpretation and processing), **accuracy** (due to the absence of key data, derived values (such as "start of service") may be factually incorrect, even though technically valid within the limited data scope) and **consistency** (discrepancy between data available and presented in the application and the DWH with different output due to missing records) as well as multiple dimensions from ISO/IEC 25012:2008 data quality standard.

4.5 Example 5: Data unreliability (violated dimension: accuracy)

The analysis determined that the data indicating a certain user characteristic was not aligned with the state in the application. For some people it was correct, but for others it was not. The error occurred due to a historical problem during data migration, and it is not possible to precisely determine which data and time periods were affected. Due to the unknown scope and the impossibility of correction within the deadlines defined by the project scope, the data was excluded from further analysis and modelling. The decision was documented and confirmed with the system users. The data was detected mainly through the analysis of data obtained through SQL queries and comparison with the state in the application and resolved by validation with the system users and documentation of business rules and exceptions.

Violated dimensions include **accuracy** (the data does not correctly reflect the real-world situation, especially for a subset of users), **completeness** (due to the unknown scope of the issue and missing context (e.g., historical change logs), the dataset is incomplete) and **consistency** (the same attribute has conflicting values between the DWH and the source application, showing inconsistent behaviour across systems) as well as multiple dimensions from ISO/IEC 25012:2008 data quality standard. It is, also important to note that in this example another violated, and very important attribute is credibility. It has been compromised because users lose trust in the data due to known inconsistencies and the lack of a way to trace or fix them.

4.6 Additional examples of data quality challenges in distributed information systems

The full list of all the possible challenges divided into a standardised guidebook would consist of countless examples.

However, here are some additional examples of data quality challenges: (1) Inconsistent User Identifiers Across Systems (primary violation: **consistency**); (2) Inconsistent User Attribute Data (e.g., Address) (primary violation: **accuracy**); (3) Unsynchronized Timing of Data Updates (primary violation: **timeliness**); (4) Inconsistent User Status Across Systems (primary violation: **consistency**) and many more.

To conclude this chapter, the mistakes in the data are omnipresent and unavoidable. One of the most important questions is how to deal with the newly recognized problem. From just a few real-life cases described above it is obvious that there are many ways of handling these kinds of situations and sometimes it is difficult to choose the best one. Nonetheless, the best way to be prepared for such non-program situations is the deep knowledge about the data quality frameworks

and standards to be able to accurately decide what amount of data loss is acceptable etc.

5 Discussion

Identified problems in data quality point to several causes, both technical and organizational. Technical causes include: **(1) Outdated IT infrastructure and disconnection between application modules** – certain data is in one application, others in another and are not connected, and even after bringing all the data into the DWH system, connection sometimes isn't possible. **(2) Historical migration errors** – incomplete control over filling processes and insufficient documentation of the transfer and control process itself makes it impossible to determine the correctness of the data.

Whilst organizational causes include: **(1) Lack of clear responsibility** – it is not clearly defined who is responsible for the accuracy and maintenance of certain data within the system, so there is often "passing the ball" instead of solving the problem. **(2) Poor coordination between IT and business users** – this leads to the situation that specific rules are not clearly defined and documented, and a lot of time is wasted only on "discovering" rules for an individual data. **(3) External system maintenance** – various external companies maintain individual parts of the system, which makes it more difficult to determine the root of the problem, as communication with each of them is required, and leads to slow corrections.

Data quality greatly affects the reliability of reporting systems. Incomplete data slows down the decision-making process due to additional checks and increases operational costs. Incomplete or inaccurate data in a fraud detection model can lead to false positives or false negatives, which reduces the efficiency of the system and undermines trust.

To improve data quality in distributed systems, organizations should establish clear ownership of data across departments, coupled with transparent governance policies that define responsibilities and escalation paths.

6 Future Work

There are countless ways to add on this ever-relevant problem in theoretical as well as pragmatical approach. Possible future discussions over the topic could encompass the quantitative case studies showing how frequent are specific problems in data quality across different distributed information systems. This approach would involve usage of metrics and statistical analysis to unbiasedly access individual challenges.

Another useful approach lies in the be development of the software for automatic inconsistency detection. This software implies creating tools that can identify data conflicts, violations of rules etc. without manual intervention. This is particularly important in environments where data is collected, stored, and processed across multiple platforms or databases because in distributed environments it is that much harder to revert and nullify the errors made in the previous phases of, for example, the ETL process.

Other useful improvements are standardization across systems in the means of creating and implementing interoperable standards and schemas to ensure consistent data formats and data definitions across different systems and organizations. Decentralized data governance models with the emphasis on how can blockchain or other distributed ledgers technologies we used to support trust, accountability, and auditability in data quality management.

7 Conclusion

Challenges with data quality to meet the appropriate standard are not anything new or previously undiscussed, they are present and will be present as long as the data itself is. However, in the data-driven world we know today, with the ever-growing need for the new data the problem is magnified exponentially if not addressed correctly (Duarte, F., 2025). That is why it is extremely important to implement the right measures and industry best practices from the very beginning. The standards such as ISO/IEC 25012 and others play a significant role in mitigating those challenges but, the standards, without the theoretical background for comprehending them fully, fall short in the real-life project tangles.

As seen countless times throughout the past not being on top of your data and managing its quality optimally, often, results in huge money, reputation and other losses (Redman, T. C., 2016). That is why any company with an aspiration to become or stay in a position of competitive player on the market should make data quality governance its priority. In other words, data quality has become the cornerstone of a profitable business in a digital and very much globalised world of 21st century. It makes or breaks it on its way to success.

References

Amatriain, X., & Basilico, J. (2015). Recommender systems in industry: A Netflix case study. In F. Ricci, L. Rokach, & B. Shapira (Eds.), Recommender systems handbook (pp. 385–419). Springer. https://doi.org/10.1007/978-1-4899-7637-6_11

Ballou, D. P., & Pazer, H. L. (1985). Modeling data and process quality in multi-input, multi-output

information systems. Management Science, 31(2), 150–162. https://doi.org/10.1287/mnsc.31.2.150

Ballou, D. P., & Tayi, G. K. (1989). Methodology for allocating resources for data quality enhancement. Communications of the ACM, 32(3), 320–329. https://doi.org/10.1145/62065.62071

Batini, C., & Scannapieco, M. (2006). Data quality: Concepts, methodologies and techniques. Springer. https://doi.org/10.1007/3-540-33173-5

Duarte, F. (2025, April 24). Amount of data created daily (2025). Exploding Topics. https://explodingtopics.com/blog/data-generated-per-day

Fisher, C. W., & Kingma, B. R. (2001). Criticality of data quality as exemplified in two disasters. Information & Management, 39(2), 109–116. https://doi.org/10.1016/S0378-7206(01)00083-0

Gualo, F., Rodriguez, M., Verdugo, J., Caballero, I., & Piattini, M. (2021). Data quality certification using ISO/IEC 25012: Industrial experiences. Journal of Systems and Software, 176, 110938. https://doi.org/10.1016/j.jss.2021.110938

Guerra-García, C., Nikiforova, A., Jiménez, S., Perez-Gonzalez, H. G., Ramírez-Torres, M., & Ontañon-García, L. (2023). ISO/IEC 25012-based methodology for managing data quality requirements in the development of information systems: Towards data quality by design. Data & Knowledge Engineering, 145, 102152. https://doi.org/10.1016/j.datak.2023.102152

International Organization for Standardization. (2008). ISO/IEC 25012:2008 — Software engineering — Software product Quality Requirements and Evaluation (SQuaRE) — Data quality model [Standard]. ISO.

International Organization for Standardization. (2015). ISO/IEC 25024:2015 — Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Measurement of data quality [Standard]. ISO.

International Organization for Standardization. (2015). ISO 8000-8:2015 — Data quality — Part 8: Information and data quality — Concepts and measuring [Standard]. ISO.

International Organization for Standardization & International Electrotechnical Commission. (2023). ISO/IEC 25010:2023 (E) — Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Product quality model (2nd ed.) [Standard]. ISO/IEC.

International Organization for Standardization & International Electrotechnical Commission. (2024). ISO/IEC 25040:2024 — Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Quality evaluation framework (2nd ed.) [Standard]. ISO/IEC.

Kahn, B. K., Strong, D. M., & Wang, R. Y. (2002). Information quality benchmarks: Product and service performance. Communications of the ACM, 45(4), 184–192. https://doi.org/10.1145/505248.506007

Kahn, M. G., Raebel, M. A., Glanz, J. M., Riedlinger, K., Steiner, J. F., & Hamilton-Lopez, M. (2015). Transparent reporting of data quality in distributed data networks. eGEMS, 3(1), Article 1052. https://doi.org/10.13063/2327-9214.1052

Llano, M. (2022, October 2). American Express: Using big data to prevent fraud. Harvard Business School Digital Initiative. https://d3.harvard.edu/platform-digit/submission/american-express-using-big-data-to-prevent-fraud/

Mahanti, R. (2018). Data governance implementation: Critical success factors. Software Quality Professional, 20(4), 4–16.

Margara, A., Cugola, G., Felicioni, N., & Cilloni, S. (2022). A model and survey of distributed data-intensive systems. arXiv. https://arxiv.org/abs/2203.10836

Miller, R., Whelan, H., Chrubasik, M., Whittaker, D., Duncan, P., & Gregório, J. (2024). A framework for current and new data quality dimensions: An overview. Data, 9(12), 151. https://doi.org/10.3390/data9120151

Redman, T. C. (2001). Data quality: The field guide. Digital Press.

Redman, T. C. (2016, September 22). Bad data costs the U.S. $3 trillion per year. Harvard Business Review. https://hbr.org/2016/09/bad-data-costs-the-u-s-3-trillion-per-year

Umar, A., Karabatis, G., Ness, L., Horowitz, B., & Elmagarmid, A. (1999). Enterprise data quality: A pragmatic approach. Information Systems Frontiers, 1(3), 279–301. https://doi.org/10.1023/A:1010006529488

Wang, R. Y. (1998). A product perspective on total data quality management. Communications of the ACM, 41(2), 58–65. https://doi.org/10.1145/269012.269022

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. https://doi.org/10.1080/07421222.1996.11518099

Woodall, P., Parlikad, A. K., & Lebrun, L. (2012). Approaches to information quality management: State of the practice of UK asset-intensive organisations. In J. Mathew, L. Ma, A. K. Parlikad, & M. A. Tiwari (Eds.), Asset condition, information systems and decision models (pp. 1–18). Springer. https://doi.org/10.1007/978-1-4471-4392-5_1