UNIVERSITY OF TARTU
Faculty of Science and Technology
Institute of Computer Science
Data Science Curriculum

Heidi Carolina Martinsaari

Toward an Automated Data Quality Rule Detection in Data Warehouses

Master's Thesis (15 ECTS)

Supervisor: Anastasija Nikiforova, PhD

Tartu 2023

Toward an Automated Data Quality Rule Detection in Data Warehouses

Abstract:
Data is a valuable asset from which information and knowledge are derived. However, business success is not depending on the amount of data only, but also on the quality of these data. On the other hand, data quality management requires a good system and the cooperation of several parties which is time-consuming and costly. Thus, it is considered if using artificial intelligence in ensuring data quality would help to avoid human errors, complement human actions, and reduce personnel costs and the workload of data quality specialists.

The objective of this thesis is to explore the current landscape of data quality solutions to find out whether these are able to automatically detect data quality rules using machine learning methods, specialising in data warehouses. For this, a systematic review of data quality software available in the market and provided in academic publications was conducted.

It was found that most of the data quality tools are used for data cleansing and fixing, meant for domain-specific databases instead of data warehouses. Meanwhile, only a few tools were capable of detecting data quality rules, not to mention implementing this in data warehouses.

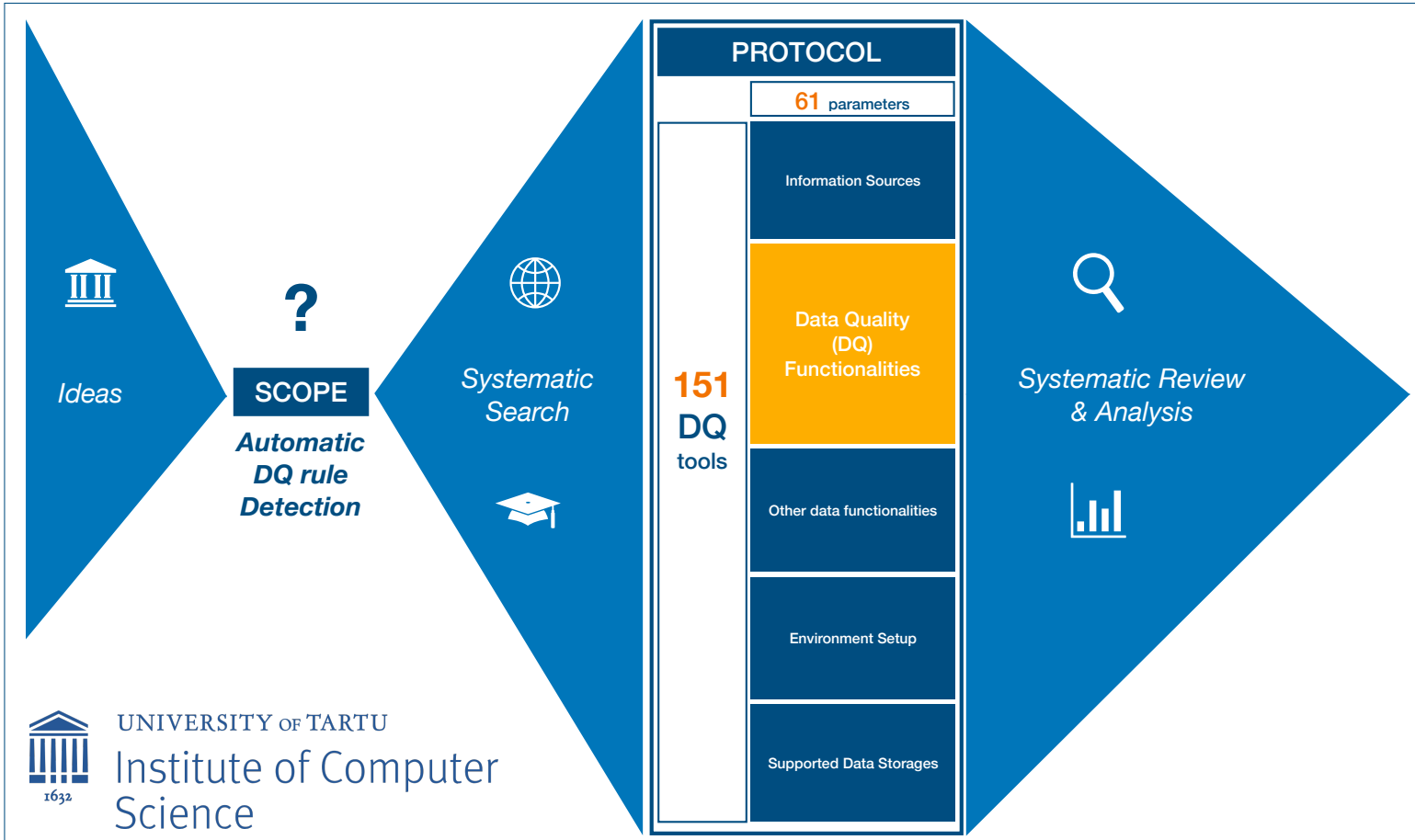
Whereas the subject of automated data quality rule detection is insufficiently covered in the academic landscape and poorly represented in the market, this thesis makes a call for action in this area.

Keywords:
Data Quality; Data Quality Rule; Data Quality Management; Data Warehouse

CERCS: P170 Computer science, numerical analysis, systems, control; P175 Informatics, systems theory; P176 Artificial Intelligence

Toward an Automated Data Quality Rule Detection in Data Warehouses

Heidi Carolina Martinsaari **Data Science** (MSc), 2023 Supervisor: Anastasija Nikiforova, PhD



UNIVERSITY OF TARTU
Institute of Computer Science

RESULTS & CONCLUSION

1/3 were **not DQ tools**

10 **DQ tools** can **detect DQ rules**.

No DQ tools have all **DW-specific DQ checks.**

75% of DQ tools meant for **data cleansing.**

73% of DQ were **unavailable to test**

Future Work

#UniTartuCS

Andmeaida andmekvaliteedi reeglite automaatse tuvastamise suunas

Lühikokkuvõte:
Andmed on väärtuslik vara, millest saab ammutada informatsiooni ja teadmisi. Sellest hoolimata, äriedu ei sõltu andmete hulgast üksnes, vaid ka andmete kvaliteedist. Teisest küljest, andmekvaliteedi tagamiseks on vaja süsteemset haldust ning mitmete spetsialistide koostööd, mis on ajamahukas ja kulukas. Seega, on mõistlik kaaluda tehisintellekti kasutamist andmekvaliteedi tagamisel, et vältida inimtekkelisi vigu, täiendada inimeste tegevusi ning vähendada personalile tehtavaid kulutusi ja andmekvaliteedi spetsialistide töömahtu.

Antud magistritöö eesmärk on uurida andmekvaliteedi tarkvarasid ja lahendusi, et välja selgitada, kas mõni lahendus on võimeline automaatselt tuvastama andmekvaliteedi reegleid, kasutades masinõppemeetodeid ja spetsialiseerudes andmeaitadele. Selleks viidi läbi süstemaatiline ülevaade turul olemasolevatest ja akadeemilises kirjanduses pakutavatest andmekvaliteedi tarkvaradest.

Töö käigus selgus, et enamus andmekvaliteedi tarkvarasid kasutatakse peamiselt andmete puhastamiseks ja parandamiseks, olles mõeldud pigem valdkonnapõhistele andmebaasidele, mitte andmeaitadele. Sealjuures, ainult vähesed rakendused olid võimelised tuvastama andmekvaliteedi reegleid, rääkimata selle rakendamisest andmeaitades.

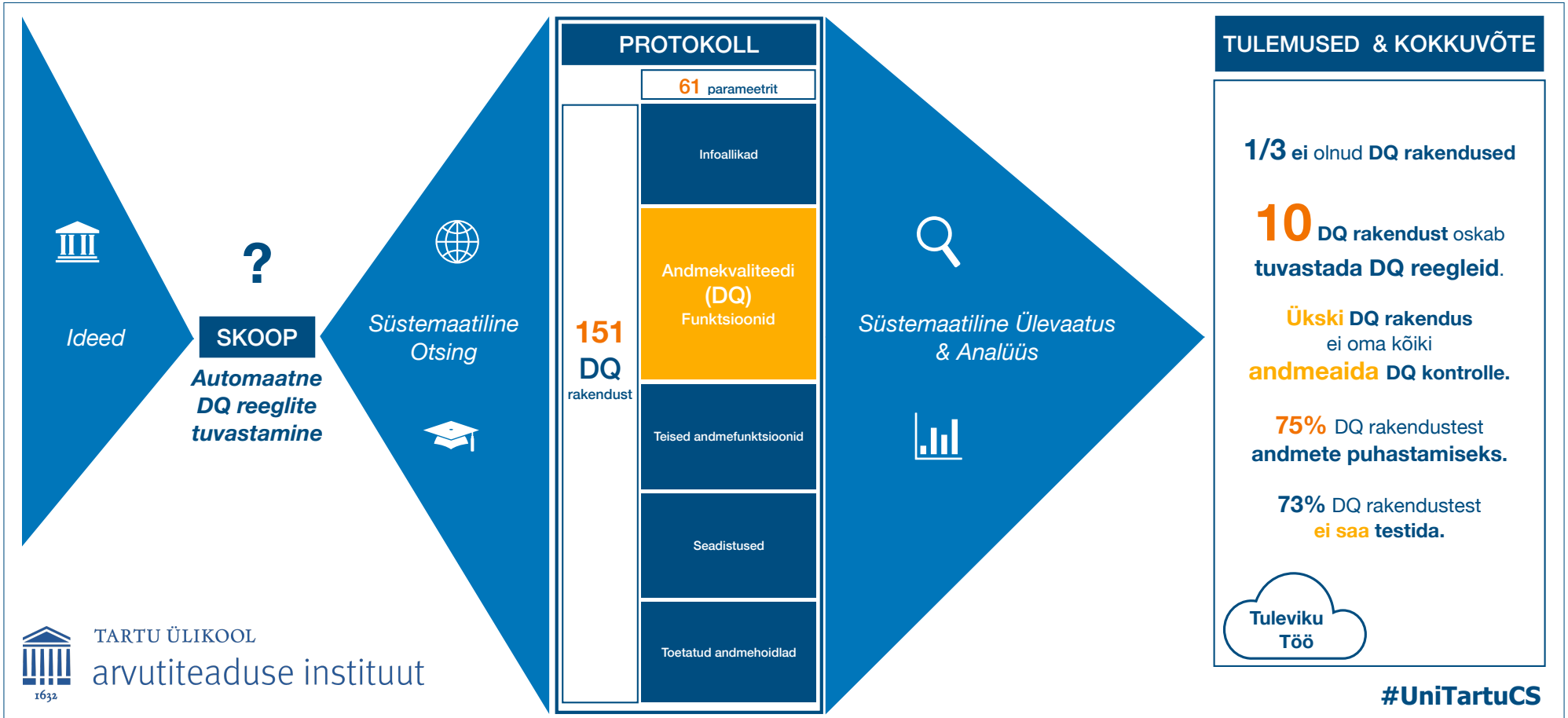
Kuna automatiseeritud andmekvaliteedi reeglite tuvastamine on ebapiisavalt kaetud akadeemilises kirjanduses kui ka väheselt esindatud tarkvaraturul, siis antud magistritöös esitatakse ka teatavad üleskutsed antud vallas.

Võtmesõnad:
Andmekvaliteet; Andmekvaliteedi reegel; Andmekvaliteedi haldamine; Andmeait

CERCS: P170 Arvutiteadus, arvutusmeetodid, süsteemid, juhtimine (automaatjuhtimisteooria); P175 Informaatika, süsteemiteooria; P176 Tehisintellekt

Andmeaida andmekvaliteedireeglite automatiseeritud tuvastamise suunas

Heidi Carolina Martinsaari **Andmeteadus** (MSc), 2023 Juhendaja: Anastasija Nikiforova, PhD



TARTU ÜLIKOOL
arvutiteaduse instituut
1632

#UniTartuCS

Contents

1 Introduction

The data is a valuable asset from which information and knowledge can be derived. Their value and credibility are directly dependent on the quality of underlying data. The data quality (DQ) topic was first raised by statistical researchers at the end of the 1960s, but it started quickly to expand at the beginning of the 1990s in the computer science field [SC02]. Computer scientists started exploring different approaches for defining, measuring, and improving the quality of electronic data stored in databases, data warehouses (DWs) and legacy systems [SC02].

The popularity of the DQ topic has even increased due to fast-growing data. International Data Corporation (IDC) predicts that the amount of data in the world will grow to 175 zettabytes by 2025 [Cou18]. While enterprises collect the data actively on a daily basis, an open issue is how to efficiently store the data, process it, and at the same time guarantee the DQ. According to [Dix20], businesses that most effectively manage DQ and optimise data storage can provide superior services to customers, improve decision-making, drive greater efficiency, and achieve assured compliance with regulators.

1.1 Enterprise Data Quality

Securing DQ is important for any enterprise. Any data-driven and profitable company may gain from data but also lose profit because of poor-quality data. According to [Dix20], a data management study by Dun & Bradstreet found that 19% of businesses had lost a customer by using inaccurate or incomplete information.

Big organisations use the DW to serve their business intelligence, reporting, analytics, and others, sourcing the data from several source databases. Such a system can be a big "maze" where manually tracing the lineage to the source and defining all needed DQ rules is complicated and time-consuming. To ensure the needed quality of the data extracted from the DW, relevant service owners are responsible to give input for DQ requirements.

While regulatory requirements force organisations to monitor their DQ effectively, any regulatory activity is not directly related to business profit [Kar22]. Whereas the costs are compared to direct incomes, DQ management tasks are relatively expensive.

The classical way to reduce personnel expenses is to automate tasks as can be seen in different industries [AR19], like the automotive industry where we can see robots building cars [ES19], and even in customer service, there are chatbots used to answer easier questions instead of asking them from employees [ZFB23]. DQ management can be similarly automated with the help of artificial intelligence.

1.2 The Motivation

The topic of this thesis is driven by the author's experience working with DQ management in a DW. The author has experience in data validation, checking DQ, analysing DQ results, including root cause analysis and tracing the data lineage, reporting DQ issues, defining business requirements jointly with DQ rules, implementing DQ rules in the DW, reporting DQ to supervisors, and therefore the author is aware of which DQ issues appear the most, how important is DQ regarding business decisions and compliance, and how much work is required for ensuring the DQ.

Therefore, it is investigated the DQ tools devoted to automated DQ rule detection which is expected to be helpful to cover the data in scope with rules more efficiently and reduce the work of data (quality) stewards. Such DQ rule detectors, perhaps involving machine learning (ML) methods, can discover checks that humans may not notice, or humans may be unaware of due to different reasons, including their level of data literacy.

In addition, the author has had several discussions with DQ professionals on the DQ and metadata quality topics. It is often discussed that data lineage functionality should be provided to DQ stewards who trace the lineage to analyse the root cause. All in all, metadata seems to be crucial from the DQ perspective as different types of DQ rules are directly related to respective information of metadata [AKT16].

1.3 An Objective of the Thesis

The objective of this thesis is to explore the DQ tools landscape and look for the tools that use the automatisation of DQ rule detection in a DW. For this, the author reviewed the existing DQ tools in the market. Also, it was familiarised with the state-of-the-art literature on both: the DQ as well as approaches for automated DQ rule definition. Finally, this thesis makes a call for action in the area. Specifically, this could involve developing a DQ tool capable of automatically detecting DQ rules in DWs and satisfying other predetermined criteria. Alternatively, further review could be conducted on the integrity constraints and other methods used for detecting DQ rules in academic literature.

1.4 Outline

This thesis consists of six sections. The introduction is followed by the second chapter which defines the concepts of DQ and DQ rules and makes an overview of the related work. The third section describes the methodology, explaining how the systematic review of DQ tools was planned and conducted. In the fourth section, the results of the review are presented towards established research questions. The fifth section summarises the results and the final section concludes the work.[1]

[1]This thesis is partly worded using *ChatGPT*, the large language model by Open AI [Ope23], and its grammar and spelling in British English are corrected with Grammarly.

2 Background and Related Work

The goal of this chapter is to provide the background knowledge for this thesis. The first subsection gives a brief theoretical overview of the DQ, and the second subsection covers the thesis-related work in the academic landscape.

2.1 Theory and Concepts

The following sections define the main concepts of this thesis, "DQ" and "DQ rule", and other related terms. There are also described the DQ lifecycle and DQ specifics in DWs.

2.1.1 Definition of Data Quality

The definition of **DQ** differs across literature, primarily drawing upon the definition of **quality** as defined by ISO 9000[2]. This standard defines quality as the extent to which the needs of the consumer are met, encompassing both the properties of the product and the intended audience or use case.

One group of definitions focus on the data consumer. For example, DQ is defined as *fit for use by data consumers* [WS96], and data is generally considered of high quality if it fits its intended uses by consumers in operations, analytics, decision-making, and planning [BS16].

Another group refers to characteristics. DQ is *a set of characteristics that data should own* [SC02], or DQ is *multidimensional measure*, where each DQ dimension or attribute indicates the certain type of DQ issue [Hae18]. All in all, there tend to be different definitions of DQ, which, in turn, affects the approaches towards its management.

The quality of data can be measured through **DQ dimensions**, which are descriptive attributes of DQ, such as completeness, timeliness, accuracy, and consistency, among others. These dimensions are highly dependent on the context, and their significance and importance can vary among organizations and data types. In addition to variations in the selection of attributes, the definition of these dimensions may also differ [CR19]. To achieve uniformity and clarity, [PC13] provides a specific set of DQ dimensions with definitions for financial institutions to adhere to:

(a) the completeness of values in the attributes that require them;

(b) the accuracy of data ensuring that the data is substantively error-free;

(c) the consistency of data ensuring that a given set of data can be matched across different data sources of the institution;

(d) the timeliness of data values ensuring that the values are up-to-date;

[2] `https://www.iso.org/obp/ui/#iso:std:iso:9000:ed-4:v1:en`

(e) the uniqueness of data ensuring that the aggregate data is free from any duplication given by filters or other transformations of source data;

(f) the validity of data ensuring that the data is founded on an adequate system of classification, rigorous enough to compel acceptance;

(g) the traceability of data, ensuring that the history, processing and location of data under consideration can be easily traced.

DQ is measured with **DQ metric** which shows the level of quality of the DQ dimensions. There are different metrics. The metric can be a subjective assessment of the quality dimension by a data consumer. Alternatively, metrics consisting of computations are measuring the DQ objectively. Some examples from Total DQ Management (TDQM) are percentages of incorrect values, an indicator of the time data was updated, a percentage of non-existent accounts and the number of records that violate referential integrity [CR19].

2.1.2 Data Quality Rules

Measuring the DQ is not mandatory for all dimensions of all data elements. According to the DQ definition "fitness for use", the DQ is user-specific and use-case-specific. Therefore, business requirements are needed for DQ rules to be specified. The DQ rule consists of two parts as followed [Plo20].

Business DQ Rule which specifies what quality means in business terms. It may also state the business process in which the rule is applied and why the rule is important to the organization.

For example, "All customers must have an identification number".

DQ Rule Specification explains what is considered “good quality” at the physical data store level.

Continuing the given example, "Customer.Identification_Nbr must not be NULL".

DQ rules can be mapped to DQ dimensions. Continuing the example of the DQ rule, it is mapped to dimension "completeness", which indicates that certain attributes should be assigned values in a data set [Los10].

Continuing the example given, there can be used different metrics to measure the DQ for this rule. Results for the example DQ rule are possible to derive with the following SQLs

```
SELECT * FROM Customer WHERE Identification_Nbr is NULL;
SELECT
    COUNT(*) Total_Count,
    SUM(CASE WHEN Identification_Nbr is NULL THEN 1) Error_Count,
    100*(Total_Count - Error_Count)/Total_Count Validity_Pct
FROM Customer
GROUP BY 1,2 HAVING Total_Count>0;
```

The initial script provides the inaccurate rows of the `Customer` table, using the records as a DQ metric. The second script returns the overall row count, the count of inaccurate rows, and the percentage of valid rows in the table. The percentage of valid rows and the count of inaccurate rows are the DQ metrics for the second script.

Deriving DQ measurements is named **data validation** or **data profiling**. Data profiling is a process of creating useful statistical summaries of data and implementing the simplest DQ rules of uniqueness, completeness, etc. [Tal23]. On the other hand, data validation is a testing process if the data conform to business rules [Inf23] that can be also mapped to DQ dimensions. Thus, there is no clear difference between these concepts. DQ measurement results are gathered into **DQ reports** or shown visually on the **dashboard** depending on the target group and quantity of the validation results.

2.1.3 Data Quality Life Cycle

The DQ life cycle, as based on the four cyclical phases of Total DQ Management [Wan98], [LO15], is as follows:

Definition phase involves analysing the data and collecting business requirements. The outcome of this phase is a logical and physical design of the information product, including attributes related to quality.

Measurement phase defines metrics for DQ and reveals problems in DQ after analysis.

Analysis phase examines the DQ problems identified in the previous phase and determines the root cause of errors.

Improvement phase entails selecting key areas for improvement, along with appropriate strategies and techniques. These strategies and techniques are implemented during the definition phase when the cycle begins anew.

That said, the DQ is implemented in a top-down direction. Business requirements are gathered from data consumers and given as input for DQ rules which can be then specified by data (quality) stewards and implemented in the data store.

2.1.4 Data Quality in Data Warehouses

The **DW** is a digital storage system that connects and harmonizes large amounts of data from many different sources. Its purpose is to "feed" business intelligence (BI), reporting, and analytics, and support regulatory requirements (see Figure 1). Thus, companies can turn their data into insight and make smart, data-driven decisions. DWs store current and historical data in one place and act as the single source of truth for an organization [SAP23].

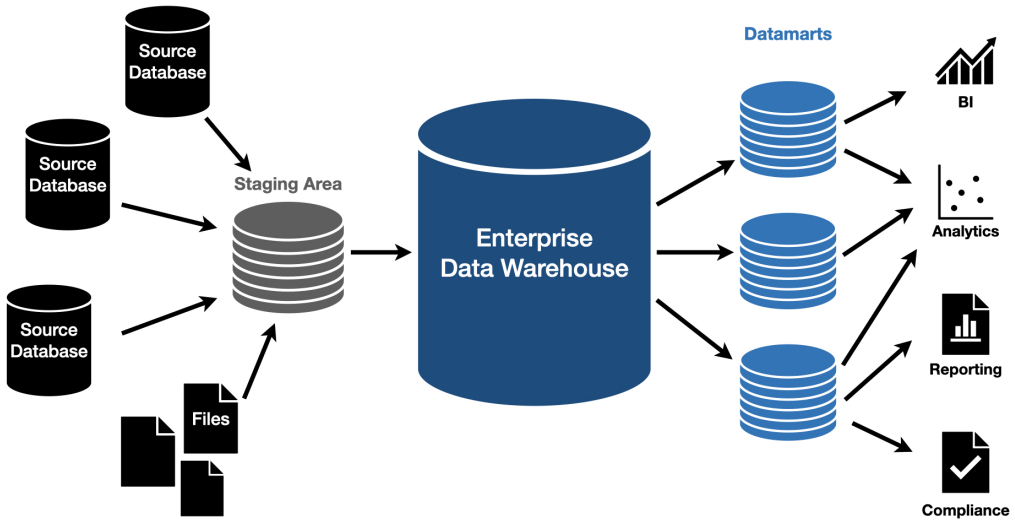


Figure 1. Architecture of DW.

As said, the purpose of the DW is to "feed" services. Therefore, it is populated only with data that is requested with business requirements and the defined information product (see Section 2.1.3). Data is sourced from source databases or files, loaded to the DW, and sent out to the data mart. Along with business requirements, all DQ rules are specified and implemented (see Section 2.1.2). There are two sources of DQ issues in DWs [LFTT19]:

1. DQ problems in data sources, i.e., poor database design, lack of accuracy in the data entry and errors entered by individuals;
2. DQ problems generated by the data integration process: loading data in and out, and calculations within the DW.

While the first type of issue is usually detected by the DQ rule of common DQ dimensions, like conformance (for checking data format), completeness or accuracy, then

the second type of issue is identified by **data reconciliation** rules. Data reconciliation is a verification phase during or after a data migration or data loading where the target data is compared against the original source data to ensure that the migration architecture has transferred the data correctly [Exp23]. Thus, data reconciliation is a kind of DQ rule or a set of rules, that helps to discover missing records, missing values, incorrect values, duplicated records, incorrectly formatted values, and broken relationships across tables or systems.

2.1.5 Metadata

Metadata is data about data [Bor03], providing comprehensive details of data, i.e., the structure, content, quality, and location [Hed16]. There are three types of metadata stored in the **data catalogue** [Edi22]:

Physical metadata covers which system data resides, the schema, table, and column or key-value level of detail;

Logical metadata provides details on how data is linked together to form larger sets. It also outlines how data flow through systems and processes, from creation to storage, transformation, and consumption. Such a path of data flow is called **data lineage**;

Conceptual metadata provides the business context for data: meaning and purpose within an enterprise, critical information about the data usage, including subject matter experts within the organization.

Metadata plays a crucial role in achieving success in data warehousing [VVS00]. As a complex system that integrates data from various sources and serves multiple purposes, the DW relies heavily on metadata. This refers to the valuable information about the data stored in the warehouse. The quality of an organisation's data is critical to its success, and the quality of the metadata is equally important to ensuring high-quality data [VVS00].

2.2 Related Work

This thesis searches for solutions for automated DQ rule detection in DWs, and surveys DQ tools' capabilities. To the best of the author's knowledge, there is no academic publication focusing on the same goal, hence this study is unique. In this section, all the work related to DQ tool surveys and automated DQ rule detection is reviewed.

2.2.1 Survey of Data Quality Tools

The goal of this section is to reflect all surveys of any DQ tools and look if there has been captured a feature of automated DQ rule detection for DWs. At first, papers were searched in Scopus using a keyword combination

S0. *("survey*" OR "review*") AND ("data quality tool*" OR "data quality software" OR "information quality tool*" OR "information quality software")*

and limiting the domain area to *Computer Science*, and language to *English*. The search resulted in eight (8) publications presented in Appendix I. All the papers were reviewed, and three (3) papers ([EW22], [HPYM18], [NLGK06]) of these were relevant, reviewing or surveying DQ tools.

Two (2) publications were surveys of domain-specific DQ tools, filling the gap between these particular domain needs and DQ software tools provided in the market. The article [NLGK06] made the survey for Realm of Engineering Asset Management in 2006, being quite old as the software landscape is fast-developing, and the paper [HPYM18] surveyed DQ tools for the clinical trial in 2018. These reviews focused on DQ tools but did not look for automated DQ rule detection.

[EW22] was a comprehensive and detailed survey of the DQ tools. In that article, there were found six hundred sixty-seven (667) tools dedicated to DQ from which only thirteen (13) tools were profoundly reviewed after implementing selection criteria. It also gave a brief overview of DQ processes that could be automated, including the automated detection of DQ rules to some extent. However, this thesis aims to look specifically for DQ tools which are able to automate the generation of DQ rules in DWs.

Furthermore, the survey mentioned earlier [EW22] cites an article on the evaluation framework for DQ tools [GNDL07]. This article centres on the functionalities of these tools, which aim to measure the quality of databases, clarifies what can be expected from such functionalities in a customer resource management (CRM) context, and suggests a general matrix for evaluating and comparing these tools. Nevertheless, this article from 2007 has become outdated due to the rapidly evolving software landscape.

2.2.2 Automated Data Quality Rule Detection

Whereas the core of the current thesis is automated DQ rule detection, the academic publications on that topic were reviewed. For searching the publications, four (4) keyword combinations were used, limiting the domain area to *Computer Science*, and language to *English* (the list of papers is provided in Appendix I):

S1. *"automated" AND "data quality" AND ("rule" OR "requirement") AND ("detection" OR "discovery" OR "recognition")*

S2. *"automat*" AND ("data quality rule" OR "data quality requirement") AND ("detection" OR "discovery" OR "recognition")*

S3. *"automat*" AND ("data quality rule*" OR "data quality requirement*") AND "data warehouse"*

S4. *"data quality" AND ("rule detect*" OR "rule discover*" OR "rule recogn*")*

Results of found papers are concluded in Table 1. The first keyword combination (S1) returned twenty-nine (29) irrelevant results. Only one (1) publication was relevant. The same article was also returned by the second (S2) and the third search (S3). The second search resulted in four (4) additional relevant publications, and the result of the third search (S3) was the same publication as the first search (S1). The fourth search (S4) resulted in five (5) additional articles and two (2) publications which were also results for the second search (S2).

Table 1. Search results: publications of the automated DQ rule detection.

Search Query	**Total**	**Irrelevant**	**Relevant**	**Publications**
S1	29	28	1	[HKO19]
S2	8	3	5	[HKO19], [YPWE11], [TS17], [MDF+15], [CIPY14], [IC15]
S3	3	2	1	[HKO19]
S4	14	7	7	[TS17], [McC08], [TH10], [SLGL16], [LWL19], [FHWX22], [CIPY14]

There were altogether ten (10) distinct publications about the automated generation of DQ rules. All of these articles were written in different years from 2008 to 2022. It is admitted that there are quite a few papers elaborating on the DQ rules discovery. Moreover, most of these papers handle the integrity constraints (ICs) as DQ rules, and if to search with keyword *"integrity rule*" AND ("detect*" OR "discover*" OR "generat*")* and limit papers by area *Computer Science* and language *English*, then twenty-five (25) additional publications from the year 1991 to 2019 appear, but which don't mention DQ rules.

[IC15] provides an overview of DQ rule discovery techniques, surveying the most commonly used ICs in the literature for detecting data inconsistencies, as well as the techniques proposed for their automatic discovery. The authors claim that automatic IC discovery is a highly useful functionality since data owners are often not DQ experts.

The newest papers [LWL19], [TS17], and [FHWX22] concentrate on discovering DQ rules in big data, focusing on optimising the detection process, using sampling the data and building efficient discovery algorithms. However, solutions are different. [TS17] predefines DQ dimensions to generate DQ rules, and [LWL19] proposes search algorithms for conditional functional dependencies (CFDs).

The latest paper [FHWX22] considers the class of entity-enhancing rules (REEs), which is a new term for more general rules comprising of conditional functional dependencies, denial constraints and matching dependencies as special cases. REEs specify rules for both entity resolution and conflict resolution and unify ML and rule-based methods. The authors propose a sampling framework with optimization strategies for REE discovery.

The paper [HKO19], which was a result of all the first three searches, presented a domain-specific language (DSL) named RADAR for specifying DQ rules, focusing on advanced DQ rules based on ARIMA models, a statistical forecasting technique. This DSL is also capable of specifying simple DQ rules, for example, a rule of uniqueness dimension. The article described how the (semi-)automated generation of DQ rules was performed based on data profiling. In addition, this rule generation is applicable to DWs and their heterogeneous sources, relational databases and document databases.

It can be seen that some academic publications have covered DQ rules detection to some extent, mostly under the topic of integrity constraints, but some of them use ML methods. However, there is no such solution that would detect DQ rules for a broad range of DQ dimensions, and at the same time specialise in DWs. Thus, this thesis will conduct a systematic review of the DQ tools provided on the market employing the methodology presented in the next section.

3 Methodology

Whereas the aim of this thesis is to investigate the landscape of possible DQ functionality automating DQ rule detection, then the essential part of the current thesis is to review the market of DQ tools. This section covers how the review of DQ tools provided in the market was planned and conducted.

3.1 Planning the Review

In this thesis, DQ tools are reviewed based on the main principles of classical systematic literature review [KB13], adapting them to surveying software tools. The following subsections elaborate on the details of the planning process.

3.1.1 Scope and Research Questions

In the planning phase of the systematic review, the first step is to establish the scope and research questions. The main goal of this thesis is to search for DQ tools that are able to discover DQ rules using ML methods. This thesis considers also solutions which can detect DQ anomalies and let users define their own DQ rules that could be based on found anomalies and for other reasons, hence acting like a semi-automated DQ rule detector. To support the scope, the author raises the following research questions:

RQ1. What are the current DQ tools proposals in the market?

Justification. This is a general question about the search results of DQ tools. It is planned to search tools from companies or magazines that provide technological reviews for software tools and academic publications to see what DQ tools there exist. The landscape of provided DQ software is described based on the initial list of tools, derived as a result of the systematic search process. It is investigated if DQ tools still exist, whether these tools are DQ management tools, meaning that these tools are dedicated to main DQ functionalities needed for DWs, like data profiling and DQ monitoring, and whether they are available to test or at least view.

RQ2. What functionalities do DQ tools have? Is there a DQ tool which can automatically detect and recommend DQ rules?

Justification. Mapping the potentially described or referenced features of available DQ tools provides an overview of their capabilities. As this thesis seeks to identify a specific type of DQ tool, this mapping provides the most significant input for selecting the appropriate tool to be included within the scope.

RQ3. Which data storages are supported? Where do the DQ tools process the organisation's data?

Justification. It is needed to know if the DQ tool is meant for DWs and does it support other sources, like files and databases, to cover also DQ controls for sources of the DW system. Additionally, it is examined whether these tools are cloud-based and the environment in which the data is processed. It is important that personal and other sensitive data is securely processed on the organisation's side because the data protection regulations prohibit the processing of personal data outside of an organisation [Uni16], and business secrets should be kept secure.

RQ4. Which methods are employed for DQ rule detection? Do the DQ tools incorporate ML techniques?

Justification. It is investigated whether DQ tools employ ML methods or similar for discovering DQ rules. Such DQ rule detectors can discover checks that humans may not notice, or humans may be unaware of due to different reasons, including their level of data literacy.

RQ5. What are the potential advantages and disadvantages that can be gleaned from existing solutions when considering the desired DQ tool solution?

Justification. During the review and analysing of the results, all the advantages are collected, and limitations are mapped to improvement ideas. Accompanying objectives are to gain an overview of the area of interest and to gather ideas for future work.

3.1.2 Searching Strategy

The second step was to establish a comprehensive list of DQ tools. For the best possible coverage of the landscape of DQ tools, it was decided to look for tools that are highly ranked by different technology reviewers like research and consulting firms, online technological publications, computer magazines, etc. Technical reviews of DQ tools were searched in Google by keyword combination *("the best data quality tools" OR "the best data quality software" OR "top data quality tools" OR "top data quality software") AND "2023"*.

Additionally, DQ tools were planned to retrieve from academic articles. The articles were searched in Scopus with two keyword combinations:

K1. *"data quality tool" OR "data quality software"*

K2. *("information quality" OR "data quality") AND ("software" OR "tool" OR "application") AND "data quality rule"*

3.1.3 Selection Criteria

For the third step, the selection criteria were set for both search directions. For searching the ranking lists of DQ tools, there were implemented following selection criteria:

- sponsored websites were excluded;
- ranking lists which were published earlier than the year 2023 were excluded;
- ranking lists in websites of other languages than English were excluded;
- websites with no technological background were excluded;
- only ranked or reviewed DQ tools were selected from not excluded ranking lists.

The selection criteria for searching the academic publications which included DQ tools were:

- publications of other fields than computer science were excluded;
- publications older than ten (10) years were excluded;
- publications not in English were excluded.

3.1.4 Review Protocol

The review process was divided into three phases which ended with excluding irrelevant tools for the next part of the review. For this, the review protocol consisted of three parts which structure with all parameters is presented in Table 2. Each parameter is complemented with an explanation and a value list.

Parameters were selected to have data to analyse and give answers towards established research questions. Parameters *ID*, *Tool Name* and *Provider* were needed for identification purposes. Thereby, *ID* is the same during the whole review. Links to *Official Website*, *Video* and *Additional Info* were meant for evidence of information. *Trialability* and *Documentation* for the main information sources and *Level of Information* as an additional decision base.

Parameters of the second phase were chosen based on the author's experience working with DQ management tasks, discussions with DQ professionals and related work [EW22]. Parameters of the DQ features were supposed to give data to gain an understanding of the DQ tools landscape. These were selected keeping in mind to main directions of ensuring DQ: DQ issue finding (DQ monitoring) and DQ issue fixing.

In the third phase, there were needed parameters to collect information on how the tools work, where they process the data and to which data sources the tools are able to connect because this thesis focuses on detecting the DQ rules in DWs but including also its sources which can be different databases, data lakes, flat files, spreadsheets, etc.

Table 2. The review protocol.

Feature	Definition	Values
Tool Name	The name of the tool/software/platform.	Name
Provider	Name of the company who provides the DQ tool.	Name
1. phase: Information Sources		
Official Website	Link to the official website of a specific tool.	Link
Video	Link to the official introductory or demo video.	Link, -
Additional Info	Link to the additional information or documentation.	Link, -
Trialability	It shows the possibility to see or try out the tool. Whether it was free to use, a free trial for a certain period, a demo version, etc.	Open-Source, a free trial, a demo, request a free trial, request a demo, not trialable, not interested
Documentation	It shows if the software company provides documentation freely or not.	Yes, No, -
Level of Information	This is the author's assessment of the level of information to continue with mapping.	Good, Partial, Low, -
Decision 1	The first decision: if to include the tool in the further review process based on exclusion criteria EC1 - EC5: EC1. Tool does not exist. EC2. Discontinued or legacy. EC3. Not a DQ tool. EC4. Part of another tool. EC5. Not enough information.	Yes, No, -
2. phase (I): DQ Management Functionalities		
Data Profiling	The tool does the data profiling and/or executes the built-in DQ rules.	Yes, No, -
Custom DQ Rules	The tool allows the user to insert custom DQ rules and execute them.	Yes, No, -
DQ Rule Definition in SQL	The tool presents the DQ rules in SQL (and allows the user to define their rules in SQL).	Yes, No, -

Continued on next page

Table 2 – continued from previous page

Feature	Definition	Values
DQ Dimensions Used	The tool classifies (or allows users to classify) DQ rules towards DQ dimensions.	Yes, No, -
DQ Rules Repository	It is possible to store DQ rules and share them with other users.	Yes, No, -
Erroneous Records Shown	Any data profiling or rule execution result is possible to drill down to relevant data records.	Yes, No, -
DQ Report Creation	It is possible to present the results of profiling or custom DQ rules in a DQ report, or whether it is possible to manually create the DQ report.	Yes, No, -
DQ Dashboard	DQ results are possible to present in dashboards.	Yes, No, -
Data Match Detection	The tool is able to detect duplicate records (exact match and fuzzy match).	Yes, No, -
Anomaly Detection	The tool is able to discover any anomalies (outliers) in the data values.	Yes, No, -
DQ Rule Detection	The tool is able to detect DQ rules.	Yes, No, -
Data Cleansing	The tool has the functionality to fix or cleanse the data.	Yes, No, -
Data Enrichment	The tool has the functionality to augment the data or fill the empty values (with reference data).	Yes, No, -
2. phase (II): Other Data Management Functionalities		
Master Data Management	The tool manages the master data.	Yes, No, -
Data Lineage	The tool is able to track the data lineage/data origination.	Yes, No, -
Data Catalogue	The tool provides data catalogue features.	Yes, No, -
Data Semantic Discovery	The tool discovers data semantics with ML methods.	Yes, No, -
Data Integration	The tool provides data integration functionalities.	Yes, No, -

Continued on next page

Table 2 – continued from previous page

Feature	Definition	Values
Scope Decision 2	The second decision: if to include the tool in the further review process based on exclusion criteria EC6 - EC8: EC6. Checks only one specific data attribute. EC7. Not detecting DQ rules or anomalies. EC8. Anomaly detection, but no DQ rules possible to define.	Yes, Yes*, No
3. phase: Environment and Connectivity Features		
Tool Environment	Location where the tool is set up.	in the cloud, hybrid, on-premises, some combination
Data Processing Environment	Location where the data is processed.	in vendor's cloud, an organisation's cloud or on-premises
API Used	The tool uses API for connecting to organisation's data store.	Yes, No, -
Flat file (.txt, .csv, .tsv)	Supported as an input.	Yes, No, -
Spreadsheet (.xlsx, .xls)	Supported as an input.	Yes, No, -
.json	Supported as an input.	Yes, No, -
Relational Database	Connection supported to relational databases.	Yes, No, -
Non-Relational Database	Connection supported to non-relational databases.	Yes, No, -
DW	Connection supported to DWs.	Yes, No, -
Data Lake	Connection supported to data lakes.	Yes, No, -
Cloud Data Storage	The tool is able to connect to the organisation's data store located in the cloud (store type in previous fields).	Yes, No, -

Continued on next page

Table 2 – continued from previous page

Feature	Definition	Values
Scope Decision 3	The third decision: if to include the tool in the further review process based on exclusion criteria EC9 - EC11: EC9. Tool is not intended for DWs. EC10. Data processing location unknown. EC11. Data processing in vendor's cloud.	Yes, Yes*, No
Criteria	Exclusion or inclusion criteria applied to the tool. Inclusion criteria: IC1. Automated DQ rule detection. IC2. Anomaly detection with custom DQ rules.	One criterion of EC1 - EC11 or IC1 - IC2.

3.1.5 Data Synthesis

For the final step, DQ tools were reviewed and data synthesised. For reviewing the tool, it was planned to

- read the information on the official website,
- read the documentation,
- download a tool or try out a platform, and review it physically,
- view a demo, and/or
- watch the official video(s).

3.2 Conducting the Review

In this section, the review process is described.

3.2.1 Executing the Search and Applying the Criteria

The DQ tools were searched with two methods: searching the ranking lists by technology reviewers and searching academic papers including DQ tools.

Sixteen (16) ranking lists were found (as of 01.04.2023) by different technology reviewers like research and consulting firms, online technological publications, computer

Table 3. Lists of DQ tools compiled by reviewers found through a Google search.

Reviewer	**Title**	**Reference**	**Nbr of Tools**
Datamation	"Best DQ Tools of 2023"	[Dat21]	11
Simplilearn	"Top DQ Tools of 2023"	[Sim23]	12
TechTarget	"7 top DQ management tools"	[Tec22b]	7
Solutions Review	"The 8 best DQ tools"	[Rev22]	8
TechRepublic	"Top DQ tools of 2022"	[Tec22a]	8
Geekflare	"The best DQ tools"	[Gee23]	8
TrustRadius	"DQ Software Overview"	[Tru23]	24
BIS (Grooper)	"The 9 Best DQ Tools 2023"	[(Gr23]	10
G2	"Best DQ Tools"	[G223]	15
Slashdot	"Best DQ Software of 2023"	[Sla23]	4
SourceForge	"DQ Software"	[Sou23]	4
PeerSpot	"Best DQ Software"	[Pee23]	10
SoftwareReviews	"Top DQ Tools"	[Sof23]	7
WebinarCare	"10 Best DQ Software for February 2023"	[Web23]	10
HubSpot	"DQ: A Comprehensive Overview"	[Hub23]	9
Gartner	"DQ Solutions Reviews and Ratings"	[Gar23]	83

magazines, etc. After applying the exclusion criteria (see Section 3.1.3) a total of one hundred twenty-eight (128) tools were retrieved from the lists listed in Table 3.

In addition, there were searched academic publications providing DQ tools with two keyword combinations (see Section 3.1.2). The first search returned fourteen (14) publications of which five (5) papers [EW22], [EGHW21], [AL20], [WOB14a], [PVA16] included thirty-five (35) DQ tools, and the second search resulted in twenty-one (21) publications of which one (1) publication [CDK$^+$22] included three (3) DQ tools. The list of research papers found is in Appendix IIb.

It was found (2) more tools as a result of the discussion with DQ experts. As certain sets of tools were mentioned more than once then altogether the author retrieved one hundred fifty-one (151) distinct DQ tools. The list of retrieved tools with relevant sources is provided in Appendix IIa.

3.2.2 Conducting the Review

The review was divided into three parts and each part ended with excluding the inappropriate tools from further review. The structure of the review, the protocol and its features are described in Table 2.

Tools were reviewed based on the available information listed in Section 3.1.5. For these tools, when it was not possible to download and test them due to the lack of even a trial version, a brief overview was conducted based on the official information on websites and videos. Also, some tools were provided with a demo with the possibility to review the tool visually but without testing it physically.

The first phase. The main aim of this phase was to map the **sources of information** of tools. The first feature of the protocol, *Official Website*, stores the link to the official page of the tool. In case it was missing, the tool was not found, and a *Criterion* "EC1. The tool does not exist." was applied to this tool.

To another group of tools was applied a *Criterion* "EC2. Legacy or discontinued." The fact of being discontinued or legacy was mostly mentioned on the official websites, or the tools were not found on their provider's official websites anymore but had any other evidence of earlier existence. It was also determined in the first phase whether the tool was a DQ tool or specialised to something else, mostly not even having any DQ features, i.e. marketing tool for handling marketing information. All non-DQ tools were applied a *Criterion* "EC3. Not a DQ tool.".

While familiarising the one with the tools based on the information on their websites, it turned out that some provided tools were actually a specific functionality of another DQ tool under investigation. These tools were applied a *Criterion* "EC4. Part of another tool investigated.".

The main expected information source was the tool itself if it was trialable. Six levels of the tool's *Trialability* were as follows: *"Open source"* for free tools, *"Free trial available"* and *"Request a free trial"* for testing the software for some time with own data, *"Demo available"* and *"Request a demo"* for looking into the tool, but not having the possibility to test it with own data sets, and *"Not available"* in case the software company or developer does not provide even the possibility to request it for try out or demo. Thereby, requests ended in not being contacted, thus, being also unavailable. Additionally, for the tools, which were excluded by criteria E1 - E4, the value was *"Not interested"*.

Another good source of the information, *Documentation* was observed if it was found or not. The documentation was searched from the official website and if not found, then there was additionally used Google search. The value list of *Documentation* consisted of *"Yes"* or *"No"* if documentation existed or not respectively, and *"-"* for the tools excluded by criteria E1 - E4.

The feature *Level of Information* was the author's assessment of the information

sufficiency based on the information on the website, in videos, documentation, and trialability. There was assessed if any DQ features were described or not. This feature had 3 main values *"Good", "Partial", "Low"* and additionally *"-"* for tools which excluded before (EC1 - EC4). Having not enough information (*"No"*) was also a stopper in the investigations and these tools were applied a *Criterion* "EC5. Not enough information."

The value of the last feature of the first phase, *Decision 1*, was *"No"* if the tool got one *Criterion* of EC1 - EC5. These tools were not included in the further review.

The second phase. In this phase, the reduced list of tools was mapped to thirteen (13) DQ management and five (5) other data management features and functionalities presented in Table 2. As mentioned before, the mapping was based on the provided sources of information. The best source of information was the downloaded tool or its trial version, but other tools which were not trialable were mapped towards functionalities which were mentioned and/or described in any other information source: the official website, video, demo, or documentation.

DQ management functionalities, including automatic DQ rule detection, and **other data management features** were reviewed to gain an understanding of the DQ tools' abilities in general. These functionalities, defined in Table 2, were reviewed to answer the questions of the second research question: What are the functionalities of DQ tools? Is there any DQ tool that can detect DQ rules? How many tools have that ability?

Based on the review results, inappropriate tools were excluded. Firstly, DQ tools which checked only one data attribute, i.e. checking validity and correctes of phone number, address, e-mail, or else, were applied a *Criterion* "EC6. Checks only one data attribute." Secondly, as it was considered alternative DQ tools, which could detect anomalies and allow defining custom DQ rules, acting like semi-automated DQ rule detection, it was applied a *Criterion* "EC7. Not detecting DQ rules or anomalies." or a *Criterion* "EC8. Anomaly detection, but no DQ rules possible to define."

All DQ tools which were excluded with Criterion (EC6 - EC8) got a value *"No"* to the protocol field *Decision 2*. DQ tools with DQ rule detection capability got a value *"Yes"* and the alternative tools that were able to detect anomalies and allowed the users to define their own rules got a value "Yes*".

The third phase. For the last phase it was reviewed the environment solution and supported data sources as this thesis is aiming to look for DQ tools which fit DWs and its sources (databases, files, data lake, etc.), and which architecture is safe and compliant with General Data Protection Regulation (GDPR) [Uni16].

For this, there were reviewed the *Tool Environment* and *Data Processing Environment*. The tool was set up in the cloud, hybrid, or on-premises, or there were provided several possibilities. Data were processed on the provider's cloud or on the organisation's side (in the cloud or on-premises). Then it was reviewed the *API Usage* for connecting data

sources, and *Supported Data Sources*, including different files and database systems from relational databases to data lakes (mentioned in Table 2), and also if the data stack could be located in the cloud.

DQ tools which did not support a DW were applied a *Criterion* "EC9. The tool is not intended for DWs." Tools for which it was not possible to determine where the vendor processes the organisation's data were excluded with *Criterion* "EC10. Data processing location unknown.", and tools that processed data on the vendor's side were excluded with a *Criterion* "EC11. Data processing in vendor's cloud."

All the tools which were applied any exclusion criterion of EC9 - EC11, got a value *"No"* in the protocol field *Decision 3*. The remaining tools got either value *"Yes"* or *"Yes*"* depending on the value of *Decision 2*. All criteria EC1 - EC11 of the tools were inserted into field *Criterion*.

The list of included DQ tools. Tools which were not applied any exclusion criteria were applied the following inclusion criteria:

IC1. Automated DQ rule detection.

IC2. Anomaly detection with custom DQ rules.

In addition to the automated DQ rule detection, it was considered an alternative solution with a combination of anomaly detection and the possibility to define custom DQ rules. This combination is equivalent to the semi-automated DQ rule generation where the machine provides data value anomalies and users can define their own rules based on these DQ issues raised.

4 Results and Analysis

In this section, the data synthesised during reviewing the DQ tools are presented by describing the results towards each established research question.

4.1 Research Question 1: The Data Quality Tools Landscape

The first research question was about the DQ tool proposals on the market. A question was raised if DQ tools still exist, whether these tools are dedicated to main DQ functionalities, and whether these are available to test or at least view.

In the first subsection, irrelevant tools are excluded from the formed list by applying the exclusion criteria shown in Table 2 in the definition of attribute "Decision 1". The second subsection presents the results of the availability to try out the tools. The third subsection it is shown statistics of available documentation and in the last section, the author gives the estimation of the level of the information on which the tools were mainly surveyed. Review results for this research question are in Appendix III.

4.1.1 Initial Tool Validation

Searching the DQ tools resulted in one hundred fifty-one (151) tools. However, the initial tool validation excludes the tools from that list (counts in Figure 2). During the review, it came out that four (4) of these tools **do not exist** as they were not found on the provider's website or elsewhere. Two (2) missing tools, *Data Preparator* and *DataMentors* were mentioned in academic article [WOB14b] in 2016, one (1) tool, *Synchronos* by *Innovative Systems* was mentioned in the website of Solutions Review [Rev22] (visited 08.04.2023), and one (1), *matchIT DQ Solutions* was named in Gartner's website [Gar23] (visited 08.04.2023).

Six (6) tools were marked **legacy or discontinued** by their providers. There were *Datiris Profiler*, *Experian Pandora*, *Talend Platform for Data Management*, two (2) *Melissa Data* solutions and *DataLever* by *RedPoint Global*. The last one was not found on its provider's website anymore but others had a comment on their websites.

Nine (9) DQ tools were excluded from further analysis because of being a **part of another investigated tool**. *Rapid Data Profiling* and *Self-Service Data Preparation* are solutions in *DataRobot AI Platform* and thus looked together. *ChainSys dataZen* is a solution for *ChainSys Smart Data Platform*, *IBM InfoSphere Information Analyzer* and *IBM InfoSphere QualityStage* are part of the platform *IBM InfoSphere Information Server*, *Syniti Master Data Management* and *Syniti Data Matching* are solutions on *Syniti Knowledge Platform*, and functionalities of *Melissa Data Data Profiler*, *MatchUp*, and *Personator* were reviewed as part of *Melissa Data DQ Components for SSIS*. These tools were admitted as duplicates.

Additionally, it was found twenty-seven (27) tools which are dedicated to other functionalities and DQ functionalities are just included to some extent. There were twelve (12) data integration tools, eight (8) tools for customer management, business analysis and marketing purposes also integrating data from different CRM systems, four (4) metadata tools without any DQ management functionalities, one (1) master data management tool without DQ management, one (1) data marketplace tool which was related to DQ only for informing purposes, showing the level of correctness of certain data field or element, and one (1) tool for visualising data of location services. All of these tools were noted **"not a DQ tool"**.

After exclusions, there were left one hundred five (105) tools or platforms dedicated to DQ to a considerable extent. Further analysis is made only on these tools.

4.1.2 Tool Trialability

The tools were trialable on different levels. There were thirteen (13) open-source tools, ten (10) available free trials and five (5) available demos. These tools were immediately trialable or viewable. For twelve (12) tools were provided with a form for requesting a free trial, and forty-three (43) tools a form for requesting a demo. Thereby, requested demos were mostly calls from salespersons with a personal introduction to the tool. The requests were left without any answer, except for two (2) cases from which one (1) software provider called and one (1) free licence for a free trial was received. Twenty-two (22) DQ tools were not trialable at all and could only be bought.

As requests resulted mostly in no answer, these could be treated also as not trialable. In conclusion, altogether seventy-seven (77) tools were not available to test or even look into, which was 73.3% of the remaining one hundred five (105) DQ tools as shown in Table 4. None of these tools were excluded.

4.1.3 Documentation

Tools which were not trialable were further reviewed based on their information on the descriptions on their official website, introductory videos, demos or documentation. Many software companies offer documentation regardless of the tool availability.

Table 4. Trialability vs documentation.

	No Documentation	**Documentation**	**All DQ Tools**
Not trialable	45	32	77
Trialable	8	20	28
Total	53	52	105

As Table 4 shows, *ca* half of the tools had documentation publicly available. However, forty-five (45) unavailable tools did not have documentation which is 42.9% out of the remaining DQ tools reviewed. These tools could be investigated only based on the information provided on their websites and other available web-based materials. Thirty-two (32) tools had documentation but also could not be tried out or viewed. Only twenty-eight (28) tools could be tested or viewed and twenty (20) of these had documentation.

4.1.4 Available Information

It was observed that many companies lacked specific details on their websites, having much text for marketing purposes instead. In protocol, it was added an additional attribute *Level of information* for the author's estimation of whether the information was sufficient or not. Seventy-two (72) tools were well described, twenty-eight (28) were partially described, and five (5) tools lacked so much information that these tools were excluded from the scope as it was impossible to decide if the tool served the goal.

DQ functionalities of master data platform *Black Tiger Platform* were not clear. Only data profiling was mentioned on their website but nothing else. *DataStreams Platform* did not also cover the concrete DQ functionalities, except validation, reporting and some DQ diagnostic. It was not described if there was used DQ rules or statistics for validations and what was actually the DQ diagnosing. It was also hard to understand what DQ functionalities *OpenDQ* had. The other DQ tool *Deduplix Ixsight* mentioned only fuzzy matching models, but it was not clear what these even do. *Talend Open Data Studio* could be freely downloaded and it had documentation available but this program did not open after instalment and the documentation was very poor. In addition, this tool was not described on its provider's website.

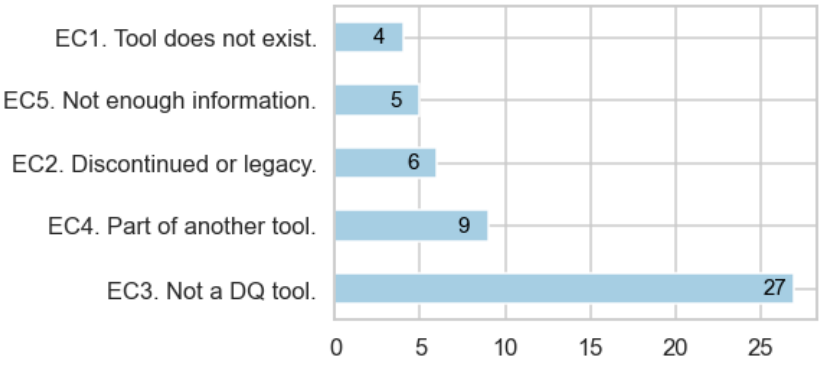


Figure 2. Counts of excluded tools in the first phase.

Altogether, in the first phase of the review, it was excluded fifty-one (51) tools which

counts are presented by exclusion criteria in Figure 2. Thus, a total of hundred (100) tools were further analysed.

4.2 Research Question 2: Features of Data Quality Tools

This section presents the observations from reviewing the remaining hundred (100) DQ tools. The second research question was about investigating which functionalities the tools have, whether there are tools that are able to automatically detect DQ rules, and how many tools are able to detect DQ rules. The review results for this research question are attached in Appendix IV.

4.2.1 Data Quality Functionalities

Each examined DQ tool was mapped to DQ functionalities which were selected for the protocol (Table 2). Out of the hundred (100) tools, there were only twelve (12) open-source tools and ten (10) available free trials which could be tested, then for the other seventy-eight (78) tools, the mapping was made if the functionality was clearly mentioned and described on the official website, documentation, demo or video.

Table 5. Relative frequencies of DQ features.

Feature	**Percentage**
Data Cleansing	75%
Data Profiling	67%
Data Enrichment	59%
Data Match Detection	55%
Custom DQ Rules	48%
Erroneous Records Shown	47%
DQ Rules Repository	41%
DQ Report Creation	35%
DQ Dashboard	35%
Anomaly Detection	26%
DQ Dimensions Used	26%
DQ Rule Detection	12%
DQ Rule Definition in SQL	6%

Relative frequencies of DQ features are presented in Table 5. It is shown that 75% of all the remaining DQ tools had a data cleansing functionality, and 67% had a data

profiling functionality. On the other hand, only 12% of DQ tools had a DQ rule detection and recommendation ability. As the goal is to focus on the DQ tools for DWs in which users are skilled in SQL which in turn gives concreteness to the DQ rules, then the author was also interested in the possibility to define the DQ rule in SQL. It seemed to be the most unpopular feature, being a feature only for 6% of DQ tools.

4.2.2 Data Functionalities

All the remaining hundred (100) DQ tools were similarly mapped to other data management functionalities: master data management, data lineage, data catalogue, data semantic discovery, and data integration. Relative frequencies of these are presented in Table 6.

Some DQ tools named in the list were single DQ solutions, i.e., *OpenRefine* or *Ataccama DQ Analyzer*. Another part was multi-functional platforms including information management features in addition to DQ management, i.e., *SAP Information Steward*, *Syniti Knowledge Platform* or *Ataccama ONE*. If the tool was only DQ management solution, then it was possible that this tool had other information management functionalities as separate solutions, i.e., *Experian Namesearch*, but these other functionalities were not mapped to these DQ tools.

Table 6. Relative frequencies of other data management functionalities.

Feature	**Percentage**
Master Data Management	30%
Data Catalogue	27%
Data Integration	25%
Data Lineage	23%
Data Semantics discovery	20%

4.2.3 Applying Exclusion Criteria

Eight (8) tools were checking only the DQ of a specific attribute, i.e., e-mail, phone, address, etc. For example *Experian Email Validation* validates e-mails online or cleanses the e-mail lists, and *Informatica Address Verification* is used to verify and validate international postal addresses in real-time in customer relationship management (CRM) systems, e-commerce sites, etc.

There were eight (8) tools, which were detecting anomalies but did not have the functionality to insert custom DQ rules by users based on anomalies discovered, i.e., *Holodetect*, *Rapid Data Profiling* and *Talend Data Fabric*. These tools are more used for

preparing the data for ML. Input data is profiled, anomalies detected and then the data will be cleansed and/or enriched.

Sixty-five (65) tools did not detect DQ rules or even anomalies. Fifty-four (54) tools of these had cleansing functionality, and thirty-one (31) of them in turn had no possibility to insert custom DQ rules. Such tools were purely cleansing tools, i.e., *Clean & Match Enterprise* by *WinPure*, *TIBCO Clarity*, *OpenRefine*, *Enlighten*, and others.

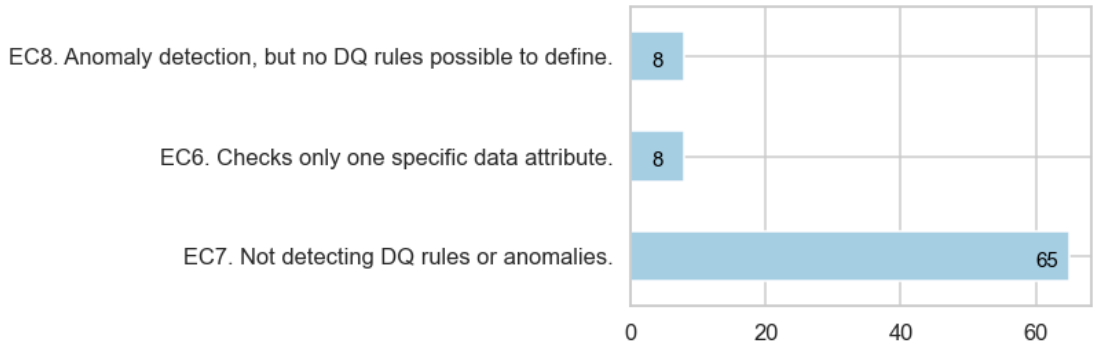


Figure 3. Counts of excluded tools in the second phase.

Exclusion criteria were applied to several tools. These counts are presented in Figure 3. Out of the remaining tools, twelve (12) DQ tools were able to detect rules and seven (7) tools were the alternative tools which could detect anomalies and allowed users to define their DQ rules. Thus, the list of DQ tools has been reduced to nineteen (19) DQ tools.

4.2.4 Clustering Data Quality Tools

To describe what DQ tools are provided in the market in the sense of what DQ and other data management functionalities these tools have, it was decided to cluster the tools by their DQ functionalities. All hundred (100) DQ tools were clustered with unsupervised ML methods *K-means* and *hierarchical clustering*.

The mappings of tools and DQ functionalities were given as input. This dataset consisted of hundred (100) rows, each for one tool, thirteen (13) columns for all DQ functionalities, and values of 1 and 0. Value 1 for the fact that the tool had the DQ functionality and value 0 for the fact that the tool did not have the DQ functionality.

The number of clusters was chosen with the *Elbow* method trying out two (2) to ten (10) clusters. The best number of clusters was $K = 2$ (Figure 4).

Clustering with methods *K-means* and *hierarchical clustering* resulted in almost the same distribution as tools were included and excluded in the second phase as shown in Figure 5. All nineteen (19) included DQ tools were clustered by K-means in one

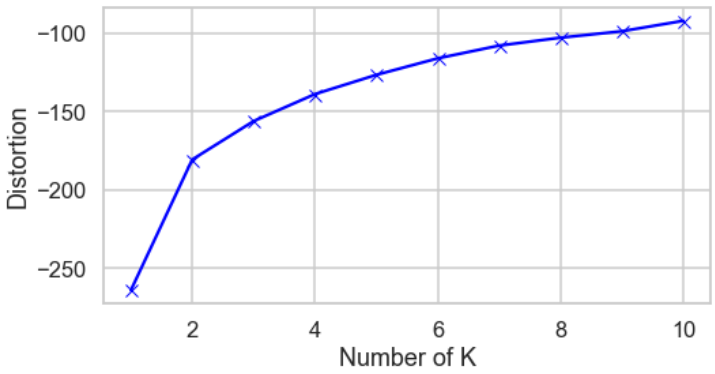


Figure 4. Elbow method for choosing the number of clusters.

group, and seventeen (17) included DQ tools by hierarchical clustering. Some excluded tools, twenty-six (26) and twenty-three (23) tools respectively were clustered under included tools (by Decision 2) but most of the excluded tools, fifty-five (55) and fifty-eight (58) tools respectively, were clustered separately matching excluded tools. In addition, clustering with *K-means* and *hierarchical clustering* differed only in five (5) tools as shown in the rightmost table in Figure 5.

Figure 5. Comparing pairwise the Decision 2 (excluded or included), *K-means* (Cluster_K2) and *hierarchical clustering* (Cluster_h2) for two clusters.

4.2.5 Included Data Quality Tools

The remaining nineteen (19) DQ tools conform to the main criterion (IC1) of this thesis or the alternative criterion (IC2).

IC1. Tools that can detect DQ rules.

IC2. Tools that can detect anomalies and allow users to define custom DQ rules.

Functionalities Relative frequencies of DQ features for tools that conform to IC1 and IC2 are shown in Table 7. It can be noticed that for IC1 defining custom rules appears for all tools. It means that a tool that is able to detect and recommend DQ rules, but also allows users to define their own rules. Rules repository is also a feature for most of the tools. Some tools might miss that functionality (or any other functionality) in the protocol because it was not mentioned on the website, documentation, etc.

Table 7. Relative frequencies of DQ features of the included tools by inclusion criteria IC1, IC2.

Feature	**IC1**	**IC2**
Custom DQ Rules	100%	100%
DQ Rules Repository	91.7%	100%
Anomaly Detection	91.7%	100%
Data Profiling	100%	85.7%
Erroneous Records Shown	100%	71.4%
DQ Report Creation	91.7%	71.4%
DQ Dashboard	75%	85.7%
DQ Dimensions Used	75%	57.1%
Data Match Detection	75%	42.9%
Data Cleansing	75%	42.9%
DQ Rule Detection	100%	0%
Data Enrichment	50%	28.6%
DQ Rule Definition in SQL	8.3%	57.1%

For DQ tools which were applied IC1 the DQ rule definition in SQL is the least frequent. Yet, expressing DQ rules in SQL is inherent to many DWs. Also, SQL presents the DQ rules in concrete "sentences" that can be validated by executing them. Thus, DW users are commonly skilled in SQL and SQL is usually preferred in data warehousing.

Data enrichment and cleansing functions are mapped for statistical purposes but are not in the scope of this thesis as it aims to look for DQ tools for DWs where data cleansing and enrichment are not used locally. So, it is expected that these features appear only as suggestions, whereas data fixing in the warehouse system is allowed only by fixing the issues in source systems and then loading the correct data to the warehouse. However, 75% of DQ rule detectors have data cleansing functionality and half of them have enrichment functionality.

It can be noticed that most of the DQ features of DQ rule detectors (IC1), except "Anomaly Detection", "DQ Dashboard" and "DQ Rule in SQL", are more frequent

features than these are for tools of anomaly detectors (IC2). All other data features shown in Table 8 are also more frequent for IC1 than for IC2 which can mean that rule detectors (IC1) are more "multifunctional" than anomaly detectors (IC2).

Table 8. Relative frequencies of other data management functionalities by the inclusion criteria by IC1, IC2.

Feature	**IC1**	**IC2**
Data Semantics discovery	75%	57.1%
Data Catalogue	75%	57.1%
Data Lineage	75%	42.9%
Master Data Management	66.7%	14.3%
Data Integration	41.7%	28.6%

It can be said that both of the included tools by IC1 and IC2 are having more functions than the tools which were excluded as shown in Figure 6. It can be also noticed that excluded tools have data cleansing and data enrichment functionalities in more cases than included tools. It refers to the fact that this thesis looks for tools focusing on DQ issue finding, not on DQ issue fixing.

Trialability and Documentation Table 9 shows the distribution of the trialability and documentation by inclusion criteria IC1 and IC2. It can be noticed that 73.7% of the DQ tools were not trialable and half of these did not have documentation either. These tools were reviewed based on less formal information found on their websites and videos.

Table 9. Triability and documentation of the DQ tools.

		No Documentation	Documentation	**Total**
IC1	Not Trialable	6	3	**9**
	Trialable	2	1	**3**
	Total	**7**	**5**	**12**

IC2	Not Trialable	1	4	**5**
	Trialable	1	1	**2**
	Total	**2**	**5**	**7**

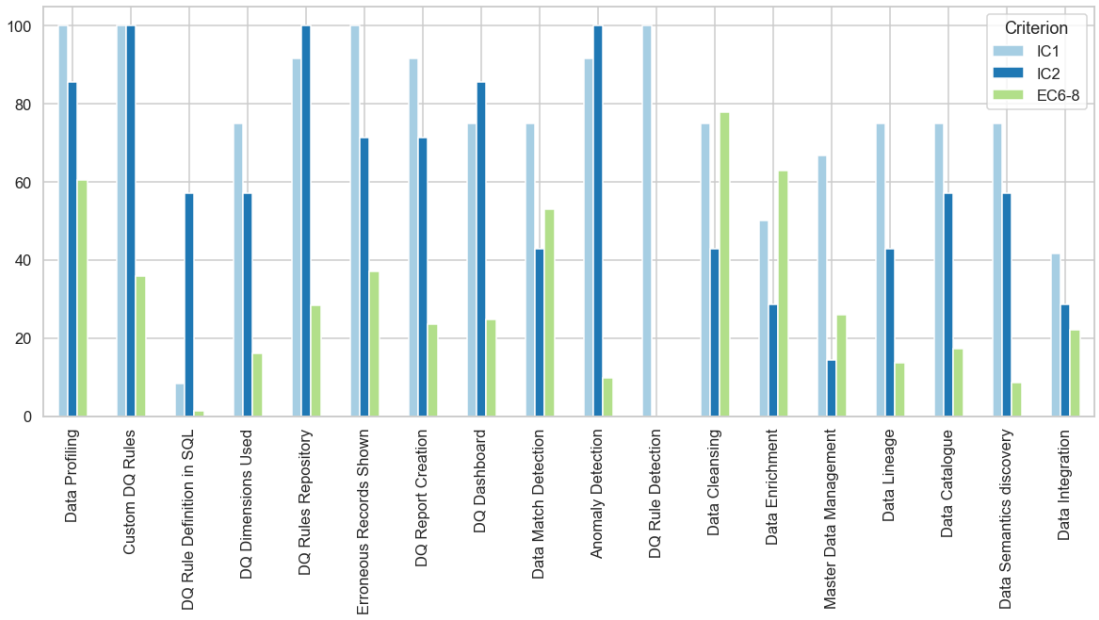


Figure 6. Relative frequencies of DQ and other data management functionalities by inclusion criteria IC1, IC2 and exclusions of the second phase.

4.3 Research Question 3: The Environment and Connectivity

This section presents the results of environment solutions and supported connections of nineteen (19) DQ tools from which twelve (12) tools were able to detect DQ rules, and seven (7) tools were alternatives, being able to detect anomalies and letting users define their own DQ rules. Results that are presented in Appendix V serve as the input for the third research question.

4.3.1 Environment and Connection Related Features

Firstly, tools were mapped to data sources to which the connection was supported, and as the current thesis focuses on DWs, then it was excluded tools which did not support the connection to DWs. There was no tool to exclude as all of these tools were supporting DWs. Another question is to which warehouse systems these adapt, and if the DQ tool is able to connect with the existing DW in the organisation, i.e., Teradata Vantage, Snowflake, Amazon Redshift, etc.

Secondly, the tools were mapped to their environment location. Most of the tools were working on the cloud. There were thirteen (13) cloud-based tools, two (2) tools were working either in the cloud or on-premises, one (1) tool was on-premises, one (1) tool was hybrid, one (1) was in the cloud or hybrid, and one (1) remained open if was

cloud-based. There were no exclusion rules for these values.

Thirdly, it was determined where tools process the data. Seven (7) tools processed the data regardless the location of it (on the vendor's side or the organisation's side), four (4) provided processing in the private cloud, one (1) on-premises, three (3) in the vendor's cloud, and four (4) did not have information for the location. As one requirement of GDPR [Uni16] is to keep the personal data inside the organisations, then tools, which data processing location was unknown or in the vendor's cloud, were excluded. The remaining list of tools was reduced to twelve (12) tools consisting of ten (10) tools that were able to detect DQ rules and two (2) additional tools which could detect anomalies and let users define custom rules.

4.3.2 Summary of Review Phases

All in all, the review process was divided into three phases where each phase exclusion criteria were applied to DQ tools as shown in Figure 7. The search process resulted in one hundred fifty-one (151) tools to which exclusion criteria EC1 - EC5 were applied. Hundred (100) tools remained. These tools were DQ tools in terms of this thesis. It is expected the DQ tool has DQ functionalities that have been brought out in the report presented in Table 2.

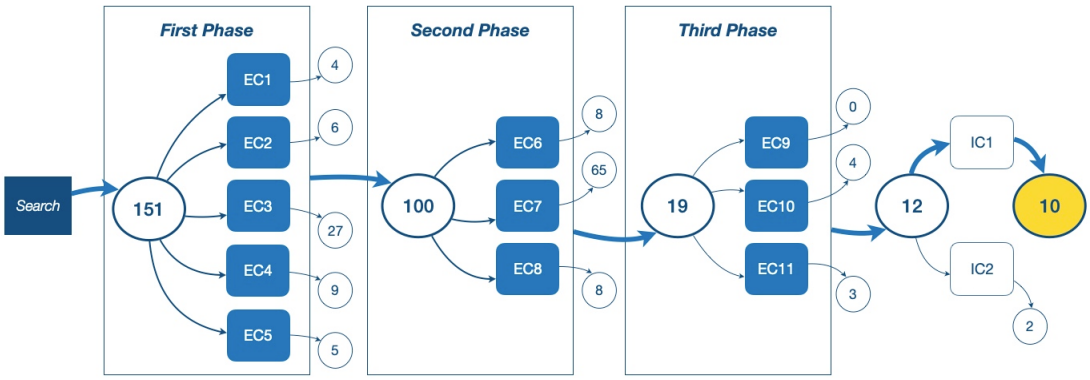


Figure 7. Review process consisted of three phases. In each phase a set of exclusion criteria were applied to DQ tools. In the final step, it is focused only on one type of included (IC1) DQ tools.

In the second phase were excluded DQ tools which did not meet the expectations of this thesis. Different types of inappropriate tools were distinguished by exclusion criteria EC6 - EC8. Applying these criteria resulted in nineteen (19) DQ tools which consisted of two (2) types of included tools, ones which could detect DQ rules and alternative tools which could detect anomalies and let users define custom DQ rules.

In the third phase, the environment solution and connectivity were reviewed for the DQ tools with desired functionalities. Tools which did not meet the expectations of this study were applied exclusion criteria E9 - E11 and there remained twelve (12) DQ tools, including ten (10) tools of the main goal and two (2) alternative solutions. Alternative tools were included in that phase for analysis but in the final step, only the main goal was reviewed because of the DQ rule detection.

4.3.3 Trialability and Documentation

58.3% of the remaining twelve (18) tools were not trialable or it was not possible to look into them as shown in Table 10. Not trialable tools consisted of three (3) tools totally unavailable to try out, four (4) tools for which it could be requested a demo, and one (1) tool for which was possible to request a free trial. For the last tool the licence key was provided for a trial, and it had documentation also*. Thus, it changed counts in Table 10.

Table 10. Distribution of trialability and documentation.

	No Documentation	Documentation	**Total**
Not Trialable	5	3(-1)*	**7**
Trialable	3	1(+1)*	**5**
Total	**8**	**4**	**12**

Thus, trialable tools consisted of one (1) trial which was requested and received (*LiTech* is also for Windows) which was the alternative solution, not being able to detect DQ rules. In addition, there were four (4) available demos. Two (2) demos were introductory videos of the tools (*Ataccama ONE*, *Anomalo*), one (1) demo (*Informatica Master Data Management*) was interactive guidance in the tool, and one (1) demo (*Collibra*) was limited platform access without being possible to test it with own data. Nevertheless, it was possible to assess tools visually.

The heatmap in Figure 8 presents the counts of tools by selection criteria and trialability. It is shown that most of the open-source tools are excluded with different criteria: eleven (11) open-source tools were excluded in the second phase because of not detecting DQ rules or anomalies (EC7) or, in case detecting anomalies, not letting define DQ rules (EC8). All the tools with the available free trial were also excluded in the second phase because of not detecting DQ rules or anomalies.

The best option to view DQ tools which were able to detect or generate DQ rules (IC1) was a demo for three (3) tools: *Ataccama ONE*, *Informatica Master Data Management* and *Collibra*. Provided demos were a video, interactive guidance and a demo tool respectively. It was not possible to test this demo tool with its own data. Four (4) DQ tools provided a form for requesting a demo, which was not received, and three (3) tools did not provide anything.

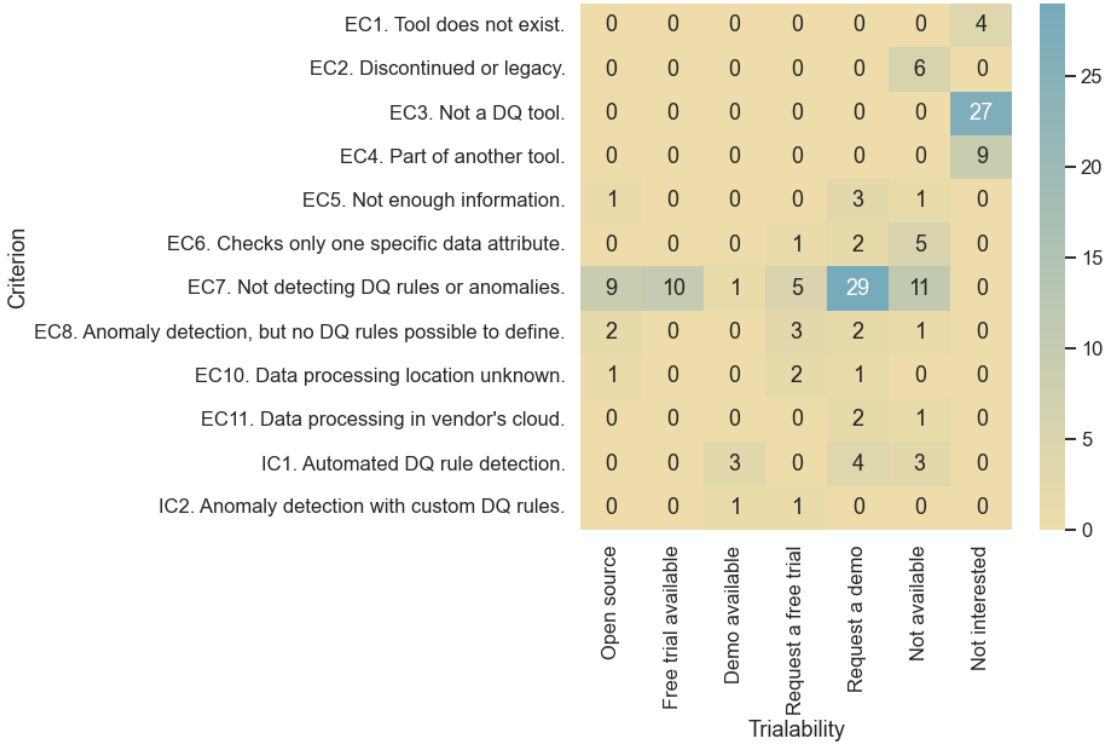


Figure 8. Heatmap of tool counts by selection criteria and trialability to understand which tools were most available.

4.3.4 Alternative Solutions

The alternative DQ tools can be taken as semi-automated DQ rule detection. It detects anomalies and lets users define their own DQ rules. Seven (7) such DQ tools were found, these all were cloud-based solutions but for four (4) tools it was not clear where the data is processed and for one (1) tool the data was processed on the vendor's cloud. Thus, after excluding these inappropriate tools there remained only two (2) suitable solutions which solutions are shortly described as follows.

Anomalo uses ML, specifically unsupervised ML, to detect DQ issues without the need to create DQ rules or set thresholds in DWs[3]. Users are also able to modify the monitoring process without using a code. Anomalo claims that classical outlier detection and time series analysis does not work as well as their unsupervised ML methods.

LiTech DQ Management assembles data validations into place, consisting DQ rule

[3]https://www.anomalo.com/post/unsupervised-data-monitoring

repository and DQ reports, and uses ML to create DQ validations on its own, including anomaly detection with alerting system[4].

Another half of the anomaly detectors, specifically eight (8) tools out of fifteen (15) DQ tools, are designed for data preparation for ML or business analytics, including also cleansing and enrichment functionalities but not using any DQ rules, i.e., *Experian DataArc360*, *Rapid Data Profiling*, and *Talend Data Fabric*.

4.4 Research Question 4: Solutions supporting the Data Quality Rule Detection

This section summarises how the automated DQ rule detection has been solved. Ten (10) DQ tools, which were able to detect DQ rules for DWs, are described based on available material and presented in Appendix VI. There is given (a) a general description presenting their features and the environment and connectivity solution, and (b) an overview of how the automated DQ rule detection has been solved.

It was observed that these DQ tools use four (4) main methods to discover DQ rules:

- using only metadata (*DQLabs Platform*),
- using built-in rules and ML (*Ataccama ONE Platform*, *DvSum*),
- using metadata and ML (*AbInitio Enterprise Data Platform*, *Informatica* products), and
- using only ML (*Collibra*, *Syniti Knowledge Platform*).

Five (5) tools out of ten (10) directly emphasise that they discover rules based on metadata and six (6) tools claims to use ML for DQ rule detection. One of the tools, *Global IDs DEEP Platform*, does not describe much about how they have solved their DQ rule detection but also includes metadata management and emphasises the importance of data lineage which is one type of metadata. Therefore, it can be admitted that metadata is one important basis for creating DQ rules by machine.

Based on the author's experience analysing the root cause of DQ issues, it is needed to understand the format of data, data lineage, relationships between data objects, etc. The author had also a discussion with other data stewards who admitted that metadata is crucial for ensuring the quality of data. In addition, [AKT16] brings out in detail how and which metadata is used to create DQ rules of specific DQ dimensions, and proposes a DQ methodology, considering both content and database metadata.

[4]https://litech.app/

On the other hand, we can see that there are some solutions which use ML for detecting DQ rules or checks, but still, there is a small number of solutions which could provide DQ rule detection for DWs and their source systems.

In addition, all reviewed ten (10) DQ tools are cloud-based tools, connecting to data sources via API to be able to connect almost every type of data source, and process customer's data where ever it is: public cloud, private cloud, or virtual private cloud.

4.5 Research Question 5: Advantages and Disadvantages of Current Solutions

All the advantages and disadvantages of existing solutions can be turned into future work ideas. The main features of existing ten (10) DQ tools, which were able to detect DQ rules, are presented in Figure 9.

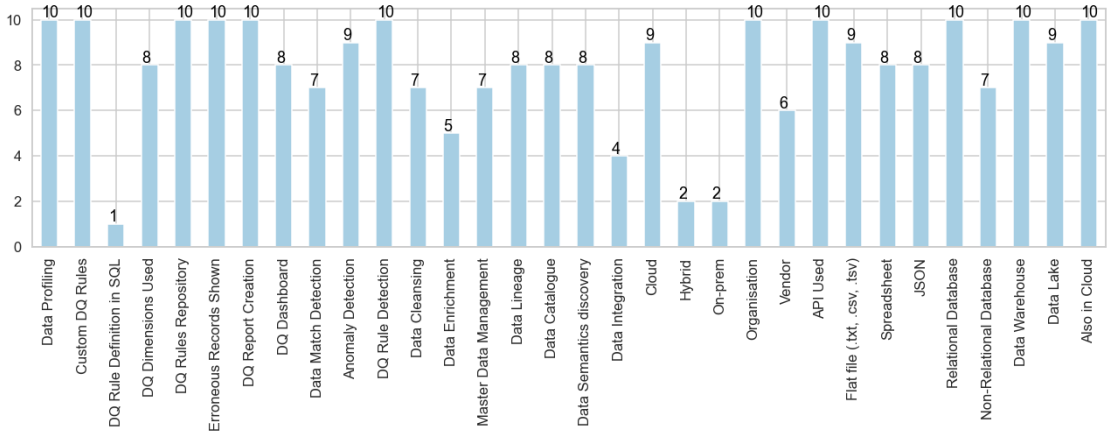


Figure 9. Frequencies of DQ, other data management, environment and connectivity features.

It is shown that all of these tools can also do data profiling, let define custom DQ rules and generate DQ reports, provide a DQ rule repository, and show erroneous records for DQ issues found. All of these process data on the organisation's realm, use API for connecting data storage, support DWs and can connect data storage in the cloud. Thereby, mentioned features, including the main feature of detecting DQ rules, are also the features of the tool expectation in the scope of this thesis. These features have been listed as advantages in Table 11.

Existing solutions have also features which do not meet the expectations set by the goal. These are brought out in Table 11 as disadvantages. DQ tool providers did not mention detecting reconciliation rules. It also seemed that the DQ rules detected were not covered in a broad range of DQ dimensions and tagging the DQ dimension to rule was

Table 11. Advantages and disadvantages of current DQ tools that can automatically detect DQ rules.

Advantages	**Disadvantages**
Detecting DQ rules.	Not detecting reconciliation rules.
Processing data on the organisation's cloud or on-premises.	Not designed for all data types (integer, float, boolean, string, date, etc.).
Using cloud computing.	Not defining DQ rules in SQL.
Using API to connect the data stack.	Option to accept, edit or reject suggested DQ rules.
Detected rules provided with erroneous data records.	Not detecting DQ rules towards different DQ dimensions, i.e., referential integrity or external consistency.
Possibility to define custom DQ rules.	Not tagging DQ rules with DQ dimension.
Proposed DQ rules can be edited, accepted and rejected.	
Tagging DQ rules with relevant data elements and business terms.	

also not noticed in many cases. Defining DQ rules in SQL was a very rare feature, but definitely necessary for rule validation purposes. It remained also unclear whether the DQ rules were suggested for sufficient list of data types. It is also important that suggested DQ rules can be handled by a data (quality) steward before implementing it, being able to modify, accept and reject the recommended DQ rules. In addition to these, it is expected that the DQ tool is working efficiently, not wasting the organisation's resources which can be solved using cloud computing benefits, like scalability and flexibility.

As seen in the Table 11, existing solutions have already many desired features but still there is much space for development.

5 Summary and Discussion

In this section, the primary findings are summarised, aiming to address the research questions. The encountered limitations are discussed, and suggestions for future work are presented.

RQ1. What are the current DQ tools proposals in the market? Technical reviewers, companies, magazines, and others provided one hundred fifty-one (151) tools as popular DQ tools for the year 2023. However, upon closer examination, it was discovered that several tools either did not exist, were outdated legacies, or had been discontinued. Additionally, certain tools overlapped with others, essentially constituting a part of another reviewed tool, resulting in duplication. After excluding these tools, a total of one hundred and five (105) DQ tools remained from the original one hundred fifty-one (151) tools obtained from diverse sources (detailed in Appendix IIa).

Additionally, there were excluded tools which lacked information being impossible to map any features of these. As a result, one hundred (100) DQ tools remained. In the next phase eight (8) additional DQ tools, which were checking only one data attribute using patterns and reference data, i.e., e-mail, phone number, address, etc., were also excluded. Thus, ninety-two (92) DQ tools were having a sufficient number of DQ features.

For comparison, another survey of DQ tools [EW22] found 667 tools from different sources, and after six (6) exclusion criteria there were left only seventeen (17) DQ tools. Half (50.82%) of the tools were excluded because these were either dedicated to specific types of data or built to measure the DQ of a proprietary tool. 16.67% of the tools focused on data cleansing without a proper DQ measurement strategy (i.e., measurements are used to modify the data, but no comprehensive reports are provided). Data profiling was provided to some extent, any of the tools had a sufficient list of DQ dimensions that have been mentioned in research papers.

The availability of free versions for testing purposes within the DQ tools market is quite limited. Out of the one hundred and five (105) DQ tools that passed the initial tool validation, only eighty-one (81) were accessible for testing. Regrettably, none of the ten (10) DQ tools encompassed by this thesis could be examined. These tools were not open-source and did not offer any free trial options, as depicted in Figure 8.

Besides, websites contain poor information, or the information there is useless, containing only exclamations for marketing purposes, i.e., emphasising how important is DQ and how big losses organisations have because of poor DQ. Only customers willing to buy the tool are contacted, and only for potential customers, the tools are presented.

RQ2. What functionalities do DQ tools have? Is there a DQ tool which can automatically detect and recommend DQ rules? After applying the first exclusion criteria (the tool does not exist, not a DQ tool, discontinued or legacy, not enough

information) there remained a list of hundred (100) tools out of one hundred fifty-one (151) tools, having a sufficient amount of DQ functionalities, and that can be called "DQ tools" as defined within this thesis. The second part of the review results is attached to Appendix III.

The most popular feature was the "Data Cleansing" functionality, appearing for 75% of all tools. It refers to the fact that most of the DQ tools are meant for fixing data issues. DQ monitoring, included in "Data Profiling", "DQ Rules (Custom or Detected)", "DQ Reports, "DQ Dashboard", is covered less. This refers to the limitation that most of the tools are not designed for DWs whereas the data is not fixed locally in the warehouses.

Furthermore, the DQ tools were found to lack other specific features of DWs. The tool designed for monitoring the DQ of the DW exhibited notable differences compared to other DQ tools, which primarily focus on data preparation for ML or business analysis. These tools predominantly address data cleansing, enrichment, and the detection of DQ issues, such as empty or irrelevant values, outliers, and inconsistent values, while the DQ tool for monitoring DQ in DWs, needs to cover only issue finding, often using DQ rules for different dimensions and based on business rules, and finally ensuring compliance of organisation to different regulations.

The focus of this thesis was the DQ rule detection in DWs. There were found only twelve (12) tools which were able to automatically detect DQ rules or checks. In addition, *Precisely Spectrum Quality* and *Information Steward* were excluded because of unknown data processing location and data processing on the vendor's cloud respectively. That lead to ten (10) DQ tools which met all the expected criteria of this thesis. It is only 10% of all DQ tools in terms of this thesis and 6.6% out of all found tools.

RQ3. Which data storages are supported? Where do the DQ tools process the organisation's data? DQ software which was in the scope of this thesis had to be able to connect to DWs and process the data in the organisation's private cloud or on-premises due to the data privacy restrictions [Uni16]. Even if all the DQ tools in scope were able to connect to DWs, these did not involve all the features of warehouses, i.e., data reconciliation rules described in Section 2.1.4.

RQ4. Which methods are employed for DQ rule detection? Do the DQ tools incorporate ML techniques? There was no exact description of how the solutions were built for automated recognition by the seven (7) providers of ten (10) DQ tools which were able to detect DQ rules. However, the main ideas were published for the majority. DQ rules were automatically generated with the help of metadata and ML methods or based on built-in rules together with ML methods.

Thereby, academic articles mainly spoke about DQ rules as integrity constraints which are also metadata of the data storage [FHWX22], [LWL19]. In articles, there was also presented a solution using the combination of built-in rules and ML methods

[TS17]. Moreover, [AKT16] shows how different DQ rules can be defined towards relevant metadata.

In conclusion, it refers to the fact that DQ management is strongly related to metadata management. On the other hand, metadata consists of data about data, including also the quality information of data [Hed16]. It could also be observed, that smart recommendation systems are built on statistical or ML methods.

All in all, there is no DQ tool which would perfectly match all requirements. It is not entirely clear whether the tools are designed especially for DWs. Software providers do not describe in detail how they have solved the automatic DQ rule detection, not to mention what is the basis for the recommendation. Nevertheless, a high-level understanding has been attained. Additionally, few tools provide DQ rule definitions in SQL and mapping to DQ dimensions. Also, it is not clear how efficient these tools are as these are not freely available for testing. Only requirements of showing erroneous records and cloud computation are fulfilled, and partly the option to define own DQ rules.

5.1 Limitations

The goal of this thesis was to research the DQ tools landscape and find out if there is any DQ tool that is able to automatically detect DQ rules. Regardless of the fact of tools' commercial or non-commercial purpose, it was needed to use the information from software providers' websites, videos and documentation if available. Even if there were used only official materials of the software providers, this kind of information is always biased and marketing-flavoured, being deficient and informal.

Another limitation was the tool's trialability. It is commonly known that commercial tools are not always available for testing. Tools that were reviewed for this thesis were either open-source tools, had demos, provided free trials, or there was needed to request a demo or a free trial. All remaining tools were totally unavailable or had no information about trying out the tool. The tools which were not trialable were impossible to test and examine fully from the consumer's perspective.

In spite of conducting a systematic search for systematic review, the risk of omitting one or another tool is always a limitation.

5.2 Future Work

Direction 1. The shortcomings identified with the existing tools and requirements defined in Section 4.5 lead to potential future work of building a DQ tool which automatically detects DQ rules and is suitable for DWs.

The reviewed tools were not purely created for DWs and these do not keep in mind the purpose of the DW and its properties. The main limitation is a missing reconciliation rule detection. Ensuring the DQ in warehouses is needed to detect both types of DQ issues in DWs as described in Section 2.1.4.

In addition, users of the DW are usually all SQL-skilled. DQ rule expression in SQL is essential for validating the rule and for later implementation. DQ rules expressed in SQL are concrete and unambiguous. On the other hand, it would be beneficial to also generate rule descriptions in natural language to involve business stakeholders in DQ work. This means involving and implementing the methods of natural language processing.

DQ rules should be detected for several data types (integer, float, boolean, string, date, character, etc.) and domains (finance, healthcare, education, etc.) to cover all data elements of a DW. In addition, a broad list of the most common DQ dimensions, as organisations tend to use different sets of DQ dimensions and some domains are even required to report DQ by DQ dimensions. For example, regulation [PC13] force financial institutions to report DQ by the specific set of DQ dimensions (completeness, accuracy, consistency, timeliness, uniqueness, validity, traceability) which were defined in Section 2.1.2.

Furthermore, [CR19] lists the frequently used DQ dimensions used by different DQ frameworks: accessibility, accuracy, an appropriate amount of data, believability, completeness, concise representation, consistency, consistent representation, currency, free-of-error, interpretability, objectivity, precision, relevancy, reputation, security, timeliness, understandability, validity, and value-added.

Current solutions were already providing recommended DQ rules with names of the related objects and fields, business terms, and potential erroneous rows. These attributes are also expected for the DQ tool of future work. The detected rules should be additionally complemented by responsible roles, like data (quality) stewards, business analysts, information owners, etc. Whereas the generation can result in an enormous amount of rules then all this information would help to filter the necessary rules by data attributes and responsible counterparties.

While using ML methods in recommendations and natural language processing for the descriptions in natural language for DQ rules, and also the metadata, then the computing can be exhaustive and it would benefit from being executed in the cloud. At the same time, the data processing should be carried out in the organisation's data stack, meaning the data would not leave the organisation's storage, to be compliant with GDPR [Uni16].

Direction 2. This thesis focused on reviewing DQ tools which can detect DQ rules. This approach holds many limitations as the software provided on the market is often not trialable and the architecture and algorithms used are hidden because of the business secret. Nevertheless, this study gave some overview of the capabilities of the DQ tools.

To gain a better understanding of what methods are used for DQ rule detection than seen in Section 4.4, it would be beneficial to make the survey of solutions brought out in academic articles. The last paper about trends of DQ rules [IC15] was published in 2015.

After that, in the last eight (8) years, there have been published additional papers [TS17], [LWL19] and [FHWX22] about automated DQ rule detection which extends this theory to the big data level. However, it seems that these theories lack similar features expected in this thesis, and for this, these should be researched.

Conclusion

The objective of this thesis was to find the DQ tools or solutions and search for tools which are able to automatically detect DQ rules in DWs. Alongside the systematic review of the DQ tools, this thesis also presented background knowledge on DQ and provided an overview of related work.

After reviewing one hundred and fifty-one (151) DQ tools and applying eleven (11) exclusion criteria, only ten (10) tools were discovered that had the capability to detect and propose DQ rules. These tools supported connecting with the DW and processing the data within the organisation's private cloud or on-premises, ensuring data confidentiality. However, it is unfortunate that certain DW-specific features were absent, such as reconciliation rules and consistency checks between attributes of different data objects.

One-third of all tools were excluded in the initial phase of the review, indicating that forty-six (46) out of the one hundred fifty-one (151) collected tools did not meet the criteria for DQ tools as defined in this thesis. Furthermore, among the remaining one hundred five (105) tools, a significant 73.3% of DQ tools were not available for trial, while 42.9% lacked both trial availability and proper documentation.

Nevertheless, a high-level understanding of how tools detect DQ rules was attained. There were mainly used metadata, built-in rules and ML methods in commercial tools. At the same time, the academic landscape, there were mainly described the usage of integrity constraints for the DQ rules, which are also a form of metadata. Regrettably, testing and experimentation of these functionalities were not feasible, and pertinent information in this regard remains unavailable.

In conclusion, the subject of automated DQ rule detection is insufficiently covered in the academic landscape and poorly represented in the market. Therefore, this thesis makes a call for action in this area. Specifically, this could involve developing a DQ tool capable of automatically detecting DQ rules in DWs and satisfying other predetermined requirements. Alternatively, further review could be conducted on the integrity constraints and other methods used for detecting DQ rules in academic literature.

Acknowledgements

I would like to express my sincere appreciation to my supervisor, Anastasija Nikiforova, for her professional guidance, supervision, and insightful responses to my inquiries. I have gained invaluable knowledge in the field of DQ, systematic (literature) review, and beyond through her mentorship.

I am grateful to all the DQ professionals who generously engaged in discussions with me and provided valuable advice and input for my thesis.

My heartfelt gratitude goes to my manager, team, and colleagues. Your unwavering support and understanding have been instrumental in keeping me motivated and determined. I am truly grateful for your invaluable assistance.

I would like to extend my deepest thanks to my family for their boundless love and support. A special acknowledgement goes to my son and daughter, who showed incredible resilience during my studies and the writing of this thesis. I love you dearly!

Lastly, I would like to express my profound gratitude to all the participants in my studies. Artjom, Dmitri, Kaja, Kertu, Mariam, Mart, Rasmus, Siim, Triin, and all others from lectures, seminars, and practical sessions with whom I engaged in discussions, collaborated in teamwork, and shared experiences. Your presence and contributions have been immensely motivating, and I am sincerely grateful for your involvement in my academic journey.

References

[AKT16] Mustafa Aljumaili, Ramin Karim, and Phillip Tretten. Metadata-based data quality assessment. *VINE Journal of Information and Knowledge Management Systems*, 46:232–250, 2016.

[AL20] O. Azeroual and W. Lewoniewski. How to inspect and measure data quality about scientific publications: Use case of wikipedia and cris databases 2020. *Algorithms*, 2020.

[AR19] Daron Acemoglu and Pascual Restrepo. The impact of automation on employment: Just the usual structural change? *Journal of Economic Perspectives*, 2019.

[Bor03] C.L. Borgman. Metadata and scholarly communication. *Annual Review of Information Science and Technology*, 37(1):3–36, 2003.

[BS16] C. Batini and M. Scannapieco. *Data and information quality*. Springer Cham, Switzerland, 1 edition, June 2016.

[CDK+22] K. Chaudhary, K. D'Spain, S. Khanal, K. Nguyen, L. Pham, G. Lall, T. Trieu, K. Kadiyala, and B. Wei. Toyota financial services data portal. In *2022 IEEE International Conference on Imaging Systems and Techniques (IST)*, pages 1–2. IEEE, June 2022.

[CIPY14] Xu Chu, Ihab F. Ilyas, Paolo Papotti, and Yin Ye. Ruleminer: Data quality rules discovery. *2014 IEEE 30th International Conference on Data Engineering*, pages 1222–1225, May 2014.

[Col] Collibra. Collibra data quality & observability: Dq rule cheat sheet. `https://www.collibra.com/us/en/resources/data-quality-rule-cheat-sheet`.

[Cou18] Tom Coughlin. 175 zettabytes by 2025. *Forbes*, 2018. `https://www.forbes.com/sites/tomcoughlin/2018/11/27/175-zettabytes-by-2025/`.

[CR19] Corinna Cichy and Stefan Rass. An overview of data quality frameworks. *IEEE Access*, 7:24634–24648, March 2019. https://ieeexplore.ieee.org/document/8642813.

[Dat21] Datamation. Best dq tools of 2023, April 2021. `https://www.datamation.com/big-data/data-quality-tools/`.

[Dix20] Mel Dixon. The cost of bad data: have you done the math? *Global Marketing Alliance*, 2020. https://www.the-gma.com/the-cost-of-bad-data-have-you-done-the-math.

[Edi22] Precisely Editor. What is a metadata and how is it used?, November 2022. https://www.precisely.com/blog/datagovernance/what-is-metadata.

[EGHW21] L. Ehrlinger, A. Gindlhumer, L.-M. Huber, and W. Wöß. 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*, 2021.

[ES19] C.P. Ezenkwu and A. Starkey. Machine autonomy: Definition, approaches, challenges and research gaps. In *Advances in Intelligent Systems and Computing*, pages 335–358. Computing Conference, 2019.

[EW22] Lisa Ehrlinger and Wolfram Wöß. A survey of data quality measurement and monitoring tools. *Frontiers in Big Data*, 5, 2022.

[Exp23] Experian. What is a data reconciliation?, 2023. https://www.experian.co.uk/business/glossary/data-reconciliation/.

[FHWX22] W. Fan, S. Han, Y. Wang, and M. Xie. Parallel rule discovery from large datasets by sampling. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 384–398. SIGMOD '22, June 2022.

[G223] G2. Best data quality tools, 2023. https://www.g2.com/categories/data-quality.

[Gar23] Gartner. The best dq tools over all, 2023. https://www.gartner.com/reviews/market/data-quality-solutions.

[Gee23] Geekflare. The best data quality tools, April 2023. https://geekflare.com/best-data-quality-tools/.

[GNDL07] V. Goasdoué, S. Nugier, D. Duquennoy, and B. Laboisse. An evaluation framework for data quality tools. In *Proceedings of the 2007 International Conference on Information Quality*. International Conference for Information Quality, January 2007.

[(Gr23] BIS (Grooper). The 9 best data quality tools 2023, January 2023. https://blog.bisok.com/general-technology/a-review-of-the-5-top-data-cleansing-tools.

[Hae18] Tom Haegemans. *Essays on Data Quality Management with Applications in the Financial Industry*. PhD thesis, KU Leuven, November 2018. https://lirias.kuleuven.be/retrieve/519416.

[Hed16] H. Hedden. *The accidental taxonomist*. Information Today, Inc., 2 edition, June 2016.

[HKO19] F. Heine, C. Kleiner, and T. Oelsner. Automated detection and monitoring of advanced data quality rules. In *Lecture Notes in Computer Science*, volume 11706, pages 238–247. Springer, Cham, August 2019.

[HPYM18] L. Houston, Y. Probst, P. Yu, and A. Martin. Exploring data quality management within clinical trials. *Applied Clinical Informatics*, 9:72–81, 2018.

[Hub23] HubSpot. Data quality: A comprehensive overview, 2023. hhttps://blog.hubspot.com/website/comprehensive-overview-of-data-quality.

[IC15] I.F. Ilyas and X. Chu. Trends in cleaning relational data: Consistency and deduplication. *Foundations and Trends in Databases*, 5:281–393, October 2015.

[Infa] Informatica. Data sheet: Cloud data quality. https://www.informatica.com/content/dam/informatica-com/en/collateral/data-sheet/cloud-data-quality_data-sheet_3688en.pdf.

[Infb] Informatica. Data sheet: Data engineering quality. https://www.informatica.com/content/dam/informatica-com/en/collateral/data-sheet/big-data-quality_data-sheet_3275en.pdf.

[Infc] Informatica. Data sheet: Multidomain mdm saas. https://www.informatica.com/content/dam/informatica-com/en/collateral/data-sheet/informatica-multidomain-mdm-saas_data-sheet_4305en.pdf.

[Inf23] Informatica. What is a data validation?, 2023. https://www.informatica.com/services-and-training/glossary-of-terms/data-validation-definition.html.

[Kar22] Soňa Karkošková. Data governance model to enhance data quality in financial institutions. *Information Systems Management*, May 2022.

[KB13] B. Kitchenham and P. Brereton. A systematic review of systematic review process research in software engineering. *Information and Software Technology*, 55(3):2049–2075, December 2013.

[LFTT19] Qi Liu, Gengzhong Feng, Giri Kumar Tayi, and Jun Tian. Managing data quality of the data warehouse: A chance-constrained programming approach. *Information Systems Frontiers*, 23:375–389, 2019.

[LO15] Daniel Linstedt and Michael Olschimke. *Building a Scalable Data Warehouse with Data Vault 2.0*. Morgan Kaufmann, 1 edition, September 2015.

[Los10] David Loshin. *The Practitioner's Guide to Data Quality Improvement*. Morgan Kaufmann, 1 edition, October 2010.

[LWL19] M. Li, H. Wang, and J. Li. Mining conditional functional dependency rules on big data. *Big Data Mining and Analytics*, 3:68–84, December 2019.

[McC08] C.J. McClanahan. Cleaning a formulation database using rule discovery techniques. *Proceedings of the 2008 International Conference on Information Quality*, 2008.

[MDF+15] Shuai Ma, Liang Duan, Wenfei Fan, Chunming Hu, and Wenguang Chen. Extending conditional dependencies with built-in predicates. *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, 27:3274–3288, December 2015.

[NLGK06] M.P. Neely, S. Lin, J. Gao, and A. Koronios. The deficiencies of current data quality tools in the realm of engineering asset management. In *12th Americas Conference on Information Systems, AMCIS 2006*, volume 1, pages 430–438. Association for Information Systems, August 2006.

[Ope23] OpenAI. Chatgpt (may 3 version), large language model, 2023. https://chat.openai.com.

[PC13] European Parliament and Council. Regulation (eu) no 575/2013 of the european parliament and of the council of 26 june 2013 on prudential requirements for credit institutions and investment firms and amending regulation (eu) no 648/2012, June 2013. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32022R0439.

[Pee23] PeerSpot. Best data quality software, March 2023. `https://www.peerspot.com/categories/data-quality`.

[Plo20] David Plotkin. *Data Stewardship*. Academic Press, 2 edition, October 2020.

[PVA16] V.S. Venkatesh Pulla, C. Varo, and M. Al. Open source data quality tools: Revisited. *Advances in Intelligent Systems and Computing*, 2016.

[Rev22] Solutions Review. The 8 best dq tools, October 2022. https://solutionsreview.com/data-management/the-best-data-quality-tools-and-software/.

[SAP23] SAP. What is a data warehouse?, 2023. https://www.sap.com/insights/what-is-a-data-warehouse.html.

[SC02] Monica Scannapieco and Tiziana Catarci. Data quality under the computer science perspective. Rome, Italy, May 2002.

[Sim23] Simplilearn. Top dq tools of 2023, January 2023. https://www.simplilearn.com/top-data-quality-tools-article.

[Sla23] Slashdot. Best data quality software of 2023, 2023. https://slashdot.org/software/data-quality/saas/?feature_data-quality=Master+Data+Management&feature_data-quality=Data+Profililng&feature_data-quality=Data+Discovery&clear.

[SLGL16] J. Sun, J. Li, H. Gao, and X. Liu. Discovery of field functional dependencies. *Proceedings - The 2015 10th International Conference on Intelligent Systems and Knowledge Engineering*, 2016.

[Sof23] SoftwareReviews. Top data quality tools, 2023. https://www.softwarereviews.com/categories/data-quality.

[Sou23] SourceForge. Data quality software, 2023. https://sourceforge.net/software/data-quality/saas/?feature_data-quality=Master+Data+Management&feature_data-quality=Data+Profililng&feature_data-quality=Data+Discovery.

[Syna] Syniti. Data sheet: Syniti knowledge platform - data quality. https://view-su2.highspot.com/viewer/6319f29dc8591c05cc343cd0?_gl=1*n3jo9h*_ga*MjY1Nzg4Njg3LjE2ODA4NDk1MDY.*_ga_D6ESBR87G0*MTY4MDg0OTUwNi4xLjEuMTY4MDg0OTcyNC41OS4wLjA.

[Synb] Syniti. Syniti knowledge platform connector and associated services. https://support.syniti.com/hc/en-us/article_attachments/9067276919319/SynitiKnowledgePlatformConnectorSecurityandRequirements.pdf.

[Tal23] Talend. What is a data profiling?, 2023. https://www.talend.com/resources/what-is-data-profiling/.

[Tec22a] TechRepublic. Top dq tools of 2022, October 2022. https://www.techrepublic.com/article/top-data-quality-tools/.

[Tec22b] TechTarget. 7 top dq management tools, September 2022. https://www.techtarget.com/searchdatamanagement/tip/Top-data-quality-management-tools.

[TH10] T.T.P. Thi and M. Helfert. Discovering dynamic integrity rules with a rules-based tool for data quality analyzing. *ACM International Conference Proceeding Series*, 2010.

[Tru23] TrustRadius. Data quality software overview, 2023. https://www.trustradius.com/data-quality#overview.

[TS17] I. Taleb and M.A. Serhani. Big data pre-processing: Closing the data quality enforcement loop. *Proceedings - 2017 IEEE 6th International Congress on Big Data*, pages 498–501, September 2017.

[Uni16] European Union. Regulation (eu) 2016/679 (gdpr), May 2016. https://eur-lex.europa.eu/eli/reg/2016/679/oj.

[VVS00] T. Vetterli, A. Vaduva, and M. Staudt. Metadata standards for data warehousing: open information model vs common warehouse metadata. *ACM Sigmod Record*, 29:68–75, 2000.

[Wan98] R. Y. Wang. A product perspective on total data quality management. *Communications of the ACM*, 41(20):58–65, February 1998.

[Web23] WebinarCare. 10 best data quality software for february 2023, 2023. https://webinarcare.com/best-data-quality-software/.

[WOB14a] P. Woodall, M. Oberhofer, and A. Borek. A classification of data quality assessment and improvement methods. *International Journal of Information Quality*, 2014.

[WOB14b] P. Woodall, M. Oberhofer, and A. Borek. A classification of data quality assessment and improvement methods. *International Journal of Information Quality*, 3(4):298–321, January 2014.

[WS96] R. Y. Wang and D. M. Strong. Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, 12(2):5–34, 1996.

[YPWE11] Peter Z. Yeh, Colin A. Puri, Mark Wagman, and Ajay K. Easo. Accelerating the discovery of data quality rules: A case study. *Proceedings of the AAAI Conference on Artificial Intelligence*, August 2011.

[ZFB23] J.J.Y. Zhanga, A. Følstadb, and C.A. Bjørklia. Organizational factors affecting successful implementation of chatbots for customer service. *JOURNAL OF INTERNET COMMERCE*, 22:122–156, 2023.

Appendix

I. List of Related Work

	Relevant	Authors	Title	Year	Source title	Citation
S0	n	Woodall P., Oberhofer M., Borek A.	A classification of data quality assessment and improvement methods	2014	International Journal of Information Quality	-
S0	**Y**	Ehrlinger L., Wöß W.	A Survey of Data Quality Measurement and Monitoring Tools	2022	Frontiers in Big Data	[EW22]
S0	n	Georgieva P., Nikolova E., Orozova D.	Data cleaning techniques in detecting tendencies in software engineering	2020	2020 43rd International Convention on Information, Communication and Electronic Technology, MIPRO 2020 - Proceedings	-
S0	n	Abdullah N., Ismail S.A., Sophiayati S., Sam S.M.	Data quality in big data: A review	2015	International Journal of Advances in Soft Computing and its Applications	-
S0	**Y**	Houston L., Probst Y., Yu P., Martin A.	Exploring data quality management within clinical trials	2018	Applied Clinical Informatics	[HPYM18]
S0	n	Ademiluyi G., Rees C.E., Sheard C.E.	Quality of smoking cessation information on the Internet: A cross-sectional survey study	2002	Journal of Documentation	-
S0	**Y**	Neely M.P., Lin S., Gao J., Koronios A.	The deficiencies of current data quality tools in the realm of engineering asset management	2006	Association for Information Systems - 12th Americas Conference On Information Systems, AMCIS 2006	[NLGK06]

S0, S1, S2, S3, S4 - search keywords presented in Section 2.2

Y - *"yes"*, n - *"no"*

Continuation of Appendix I						
	Relevant	Authors	Title	Year	Source Title	Citation
S0	n	Barata J., da Cunha P.R., Costa C.C.	The foundations for an IS quality culture in the context of ISO 9001	2013	Proceedings of the European, Mediterranean and Middle Eastern Conference on Information Systems, EMCIS 2013	-
S1	n	Borovina Josko J.M., Oikawa M.K., Ferreira J.E.	A formal taxonomy to improve data defect description	2016	Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)	-
S1	n	Hajian S., Domingo-Ferrer J.	A methodology for direct and indirect discrimination prevention in data mining	2013	IEEE Transactions on Knowledge and Data Engineering	-
S1	n	Ehrlinger L., Wöß W.	A Survey of Data Quality Measurement and Monitoring Tools	2022	Frontiers in Big Data	[EW22]
S1	n	Kropp T., Bombeck A., Lennerts K.	An Approach to Data Driven Process Discovery in the Cost Estimation Process of a Construction Company	2021	Proceedings of the International Symposium on Automation and Robotics in Construction	-
S1	n	Babu M.C., Pushpa S.	An efficient discrimination prevention and rule protection algorithms avoid direct and indirect data discrimination in web mining	2018	International Journal of Intelligent Engineering and Systems	-
S1	n	Ardeti V.A., Kolluru V.R., Varghese G.T., Patjoshi R.K.	An Outlier Detection and Feature Ranking based Ensemble Learning for ECG Analysis	2022	International Journal of Advanced Computer Science and Applications	-

S0, S1, S2, S3, S4 - search keywords presented in Section 2.2

Y - *"yes"*, n - *"no"*

Continuation of Appendix I						
	Relevant	Authors	Title	Year	Source Title	Citation
S1	n	Alferes J., Vanrolleghem P.A.	Automated data quality assessment: Dealing with faulty on-line water quality sensors	2014	Proceedings - 7th International Congress on Environmental Modelling and Software: Bold Visions for Environmental Modeling, iEMSs 2014	-
S1	**Y**	Heine F., Kleiner C., Oelsner T.	Automated Detection and Monitoring of Advanced Data Quality Rules	2019	Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)	[HKO19]
S1	n	Malik W.A., Unwin A.	Automated error detection using association rules	2011	Intelligent Data Analysis	-
S1	n	Ataeyan M., Daneshpour N.	Automated Noise Detection in a Database Based on a Combined Method	2021	Statistics, Optimization and Information Computing	-
S1	n	Redyuk S., Kaoudi Z., Markl V., Schelter S.	Automating data quality validation for dynamic data ingestion	2021	Advances in Database Technology - EDBT	-
S1	n	Gawhade R., Bohara L.R., Mathew J., Bari P.	Computerized Data-Preprocessing to Improve Data Quality	2022	ICPC2T 2022 - 2nd International Conference on Power, Control and Computing Technologies, Proceedings	-
S1	n	Freemon D.M., Lim K.-T., Becla J., Dubois-Felsman G.P., Kantor J.	Data management cyberinfrastructure for the Large Synoptic Survey Telescope	2012	Proceedings of SPIE - The International Society for Optical Engineering	-

S0, S1, S2, S3, S4 - search keywords presented in Section 2.2

Y - *"yes"*, n - *"no"*

Continuation of Appendix I						
	Relevant	Authors	Title	Year	Source Title	Citation
S1	n	Mejia F., Shyu M.-L., Nanni A.	Data quality enhancement and knowledge discovery from relevant signals in acoustic emission	2015	Mechanical Systems and Signal Processing	-
S1	n	Zhang R., Albrecht A., Kausch J., Putzer H.J., Geipel T., Halady P.	DDE process: A requirements engineering approach for machine learning in automated driving	2021	Proceedings of the IEEE International Conference on Requirements Engineering	-
S1	n	Fan W., Lu H., Madnick S.E., Cheung D.	Discovering and reconciling value conflicts for numerical data integration	2001	Information Systems	-
S1	n	Rekatsinas T., Dong X.L., Getoor L., Srivastava D.	Finding quality in quantity: The challenge of discovering valuable sources for integration	2015	CIDR 2015 - 7th Biennial Conference on Innovative Data Systems Research	-
S1	n	O'Neill P., Magoulas G.D., Liu X.	Improved processing of microarray data using image reconstruction techniques	2003	IEEE Transactions on Nanobioscience	-
S1	n	Ryazantsev O., Khoroshun G., Riazantsev A., Strelkova T.	Informational model of optical signals and images in machine vision systems	2021	Examining Optoelectronics in Machine Vision and Applications in Industry 4.0	-
S1	n	Teiken Y., Brüggemann S., Appelrath H.-J.	Interchangeable consistency constraints for public health care systems	2010	Proceedings of the ACM Symposium on Applied Computing	-
S1	n	Ahmed M., Taconet C., Ould M., Chabridon S., Bouzeghoub A.	IoT data qualification for a logistic chain traceability smart contract	2021	Sensors	-

S0, S1, S2, S3, S4 - search keywords presented in Section 2.2

Y - *"yes"*, n - *"no"*

Continuation of Appendix I						
	Relevant	Authors	Title	Year	Source Title	Citation
S1	n	Zvara Z., Szabó P.G.N., Balázs B., Benczúr A.	Optimizing distributed data stream processing by tracing	2019	Future Generation Computer Systems	-
S1	n	Silva-Ramírez E.-L., Pino-Mejías R., López-Coello M.	Single imputation with multilayer perceptron and multiple imputation combining multilayer perceptron and k-nearest neighbours for monotone patterns	2015	Applied Soft Computing	-
S1	n	Mahajan P.	Textual Data Quality at Scale for High Dimensionality Data	2022	2022 International Conference on Data Science, Agents and Artificial Intelligence, ICDSAAI 2022	-
S1	n	Rajan N.S., Gouripeddi R., Mo P., Madsen R.K., Facelli J.C.	Towards a content agnostic computable knowledge repository for data quality assessment	2019	Computer Methods and Programs in Biomedicine	-
S1	n	Wang P., He Y.	Uni-DeTecT: A unified approach to automated error detection in tables	2019	Proceedings of the ACM SIGMOD International Conference on Management of Data	-
S1	n	Poon L., Farshidi S., Li N., Zhao Z.	Unsupervised Anomaly Detection in Data Quality Control	2021	Proceedings - 2021 IEEE International Conference on Big Data, Big Data 2021	-
S1	n	Steiniger S., Taillandier P., Weibel R.	Utilising urban context recognition and machine learning to improve the generalisation of buildings	2010	International Journal of Geographical Information Science	-

S0, S1, S2, S3, S4 - search keywords presented in Section 2.2

Y - *"yes"*, n - *"no"*

Continuation of Appendix I						
	Relevant	Authors	Title	Year	Source Title	Citation
S1	n	Strazdins G., Mednis A., Zviedris R., Kanonirs G., Selavo L.	Virtual ground truth in vehicular sensing experiments: How to mark it accurately	2011	SENSORCOMM 2011 - 5th International Conference on Sensor Technologies and Applications and WSNSCM 2011, 1st International Workshop on Sensor Networks for Supply Chain Management	-
S2	**Y**	Yeh P.Z., Puri C.A., Wagman M., Easo A.K.	Accelerating the discovery of data quality rules: A case study	2011	Proceedings of the National Conference on Artificial Intelligence	[YPWE11]
S2	**Y**	Heine F., Kleiner C., Oelsner T.	Automated Detection and Monitoring of Advanced Data Quality Rules	2019	Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)	[HKO19]
S2	**Y**	Taleb I., Serhani M.A.	Big Data Pre-Processing: Closing the Data Quality Enforcement Loop	2017	Proceedings - 2017 IEEE 6th International Congress on Big Data, BigData Congress 2017	[TS17]
S2	**Y**	Ma S., Duan L., Fan W., Hu C., Chen W.	Extending Conditional Dependencies with Built-in Predicates	2015	IEEE Transactions on Knowledge and Data Engineering	[MDF+15]
S2	n	Dallachiesat M., Ebaid A., Eldawy A., Elmagarmid A., Ilyas I.F., Ouzzani M., Tang N.	NADEEF: A commodity data cleaning system	2013	Proceedings of the ACM SIGMOD International Conference on Management of Data	-

S0, S1, S2, S3, S4 - search keywords presented in Section 2.2

Y - *"yes"*, n - *"no"*

Continuation of Appendix I						
	Relevant	Authors	Title	Year	Source Title	Citation
S2	n	Abboura A., Sahri S., Baba-Hamed L., Ouziri M., Benbernou S.	Quality-based online data reconciliation	2016	ACM Transactions on Internet Technology	-
S2	**Y**	Chu X., Ilyas I.F., Papotti P., Ye Y.	RuleMiner: Data quality rules discovery	2014	Proceedings - International Conference on Data Engineering	[CIPY14]
S2	**Y**	Ilyas I.F., Chu X.	Trends in cleaning relational data: Consistency and deduplication	2015	Foundations and Trends in Databases	[IC15]
S3	n	Abdellaoui S., Bellatreche L., Nader F.	A Quality-Driven Approach for Building Heterogeneous Distributed Databases: The Case of Data Warehouses	2016	Proceedings - 2016 16th IEEE/ACM International Symposium on Cluster, Cloud, and Grid Computing, CCGrid 2016	-
S3	n	Lettner C., Stumptner R., Bokesch K.-H.	An approach on ETL attached data quality management	2014	Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)	-
S3	**Y**	Heine F., Kleiner C., Oelsner T.	Automated Detection and Monitoring of Advanced Data Quality Rules	2019	Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)	[HKO19]
S4	n	Dimitoglou G., Rotenstreich S.	A system for association rule discovery in emergency response data	2007	Innovations and Advanced Techniques in Computer and Information Sciences and Engineering	-

S0, S1, S2, S3, S4 - search keywords presented in Section 2.2

Y - *"yes"*, n - *"no"*

Continuation of Appendix I						
	Relevant	Authors	Title	Year	Source Title	Citation
S4	**Y**	Goasdoué V., Nugier S., Duquennoy D., Laboisse B.	An evaluation framework for data quality tools	2007	Proceedings of the 2007 International Conference on Information Quality, ICIQ 2007	[GNDL07]
S4	n	Abu Ahmad H., Wang H.	Automatic weighted matching rectifying rule discovery for data repairing: Can we discover effective repairing rules automatically from dirty data?	2020	VLDB Journal	-
S4	**Y**	Taleb I., Serhani M.A.	Big Data Pre-Processing: Closing the Data Quality Enforcement Loop	2017	Proceedings - 2017 IEEE 6th International Congress on Big Data, BigData Congress 2017	[TS17]
S4	**Y**	McClanahan C.J.	Cleaning a formulation database using rule discovery techniques	2008	Proceedings of the 2008 International Conference on Information Quality, ICIQ 2008	[McC08]
S4	n	van Cruchten R.M.E.	Data quality in process mining: A rule-based Approach	2019	CEUR Workshop Proceedings	-
S4	n	Fan W.	Data quality: Theory and practice	2012	Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)	-
S4	**Y**	Thi T.T.P., Helfert M.	Discovering dynamic integrity rules with a rules-based tool for data quality analyzing	2010	ACM International Conference Proceeding Series	[TH10]

S0, S1, S2, S3, S4 - search keywords presented in Section 2.2

Y - *"yes"*, n - *"no"*

Continuation of Appendix I						
	Relevant	Authors	Title	Year	Source Title	Citation
S4	**Y**	Sun J., Li J., Gao H., Liu X.	Discovery of field functional dependencies	2016	Proceedings - The 2015 10th International Conference on Intelligent Systems and Knowledge Engineering, ISKE 2015	[SLGL16]
S4	n	Gadish D.A.	Introducing elasticity for spatial knowledge management	2008	International Journal of Knowledge Management	-
S4	**Y**	Li M., Wang H., Li J.	Mining conditional functional dependency rules on big data	2020	Big Data Mining and Analytics	[LWL19]
S4	**Y**	Fan W., Han Z., Wang Y., Xie M.	Parallel Rule Discovery from Large Datasets by Sampling	2022	Proceedings of the ACM SIGMOD International Conference on Management of Data	[FHWX22]
S4	**Y**	Chu X., Ilyas I.F., Papotti P., Ye Y.	RuleMiner: Data quality rules discovery	2014	Proceedings - International Conference on Data Engineering	[CIPY14]
S4	n	Vo L.T.H., Cao J., Rahayu W., Nguyen H.-Q.	Structured content-aware discovery for improving XML data consistency	2013	Information Sciences	-

S0, S1, S2, S3, S4 - search keywords presented in Section 2.2

Y - *"yes"*, n - *"no"*

IIa. DQ Tools: Sources

ID	Tool	Academic papers	Datamation	Simplilearn	TechTarget	Solutions Review	TechRepublic	Geekflare	TrustRadius	BIS (Grooper)	G2	Slashdot	SourceForge	PeerSpot	SoftwareReviews	WebinarCare	HubSpot	Gartner	DQ Experts
1	Data Preparator	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n
2	Holodetect	**Y**	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n
3	MetricDoc	**Y**	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n
4	DataMentors	**Y**	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n
5	DQ-MeeRKat	**Y**	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n
6	MobyDQ	**Y**	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n
7	Great Expectations	**Y**	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n
8	AbInitio Enterprise Data Platform	n	n	n	n	n	n	**Y**	n	n	n	n	n	n	n	n	n	n	n
9	Acceldata	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	**Y**
10	DQ*Plus Enterprise Suite	n	n	n	n	n	n	n	**Y**	n	n	n	n	n	n	n	n	n	n
11	Amperity CDP	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	**Y**	n
12	Anomalo	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	**Y**	n
13	Apache Griffin	n	n	n	n	n	n	n	n	n	**Y**	n	n	n	n	n	n	**Y**	n
14	Aggregate Profiler	**Y**	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n
15	Attaccama DQ-Analyzer	**Y**	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n
16	Ataccama ONE	**Y**	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n
17	dspCompose	**Y**	n	n	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	n	n	n	n	n	n	**Y**	**Y**	n
18	CRM Cleaning	**Y**	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n
19	Black Tiger Platform	n	n	n	n	n	n	n	n	n	**Y**	n	n	n	n	**Y**	n	n	n
20	ChainSys dataZen	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	**Y**	n

Y - *"yes"*, n - *"no"*

Continuation of Appendix IIa

ID	Tool	Academic papers	Datamation	Simplilearn	TechTarget	Solutions Review	TechRepublic	Geekflare	TrustRadius	BIS (Grooper)	G2	Slashdot	SourceForge	PeerSpot	SoftwareReviews	WebinarCare	HubSpot	Gartner	DQ Experts
21	Smart Data Platform	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	**Y**	n
22	Claravine	n	n	n	n	n	n	n	n	n	n	n	n	n	**Y**	n	n	n	n
23	ClearAnalytics	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	**Y**	n
24	Cloudingo	n	n	n	n	n	n	n	**Y**	n	n	n	n	n	n	n	n	n	n
25	Collibra Platform	n	**Y**	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n
26	Cribl Stream	n	n	n	n	n	n	n	**Y**	n	**Y**	n	n	n	n	n	n	**Y**	n
27	CuriumDQM	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	**Y**	n
28	D&B Connect	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	**Y**	n
29	D&B Optimizer	n	n	n	n	n	n	n	n	n	**Y**	n	n	n	n	n	n	n	n
30	DataMatch Enterprise	n	n	n	n	n	n	n	n	n	n	n	n	n	n	**Y**	n	n	n
31	Datactics Self-Service Data Quality Platform	**Y**	**Y**	**Y**	n	n	**Y**	**Y**	n	n	n	n	n	n	n	n	n	**Y**	n
32	Dataedo	**Y**	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	**Y**	n
33	DataStreams Platform	n	n	n	n	n	**Y**	**Y**	n	n	n	n	n	n	n	n	n	n	n
34	Ultimate Data Export	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	**Y**	n
35	Datiris Profiler	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	**Y**	n	n
36	Dedupely	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	**Y**	n
37	MyDataQ	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	**Y**	n	n
38	DQE One	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	**Y**	n
39	DQLABS Platform	n	n	n	n	n	n	n	n	n	**Y**	n	n	n	n	n	n	n	n
40	Duco Platform	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	**Y**	n
41	DvSum	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	**Y**	n

Y - *"yes"*, n - *"no"*

Continuation of Appendix IIa

ID	Tool	Academic papers	Datamation	Simplilearn	TechTarget	Solutions Review	TechRepublic	Geekflare	TrustRadius	BIS (Grooper)	G2	Slashdot	SourceForge	PeerSpot	SoftwareReviews	WebinarCare	HubSpot	Gartner	DQ Experts
42	Edge Delta	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	**Y**	n
43	Exmon	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	**Y**	n
44	Experian Aperture Data Studio	n	n	n	n	n	n	n	n	n	**Y**	n	n	n	n	n	n	n	n
45	Experian DataArc360	n	n	**Y**	n	n	n	n	**Y**	n	n	n	n	n	n	n	n	**Y**	n
46	Experian Email Validation	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	**Y**	n
47	Experian Namesearch	n	n	n	n	n	n	n	**Y**	n	n	n	n	n	n	n	n	n	n
48	Experian Pandora (Legacy)	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	**Y**	n
49	Experian Phone Validation	**Y**	n	**Y**	n	n	n	n	n	n	n	n	n	n	n	n	n	**Y**	n
50	Experian Prospect IQ	n	n	n	n	n	n	n	**Y**	n	n	n	n	n	n	n	n	n	n
51	Flatfile	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	**Y**	n
52	Global IDs Data Quality Suites	n	n	n	n	n	n	n	n	n	**Y**	n	n	n	n	n	n	n	n
53	OpenRefine	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	**Y**	n
54	matchIT Data Quality Solutions	**Y**	**Y**	**Y**	n	n	**Y**	**Y**	n	n	n	n	n	n	n	n	**Y**	n	n
55	HERE Platform	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	**Y**	n
56	HubSpot Operations Hub	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	**Y**	n
57	DataCleaner	n	n	n	n	n	n	n	**Y**	n	n	n	n	n	n	n	**Y**	n	n
58	InfoZoom	**Y**	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n
59	ibi Data Quality	**Y**	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n
60	ibi Omni-Gen	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	**Y**	n

Y - *"yes"*, n - *"no"*

Continuation of Appendix IIa																			
ID	Tool	Academic papers	Datamation	Simplilearn	TechTarget	Solutions Review	TechRepublic	Geekflare	TrustRadius	BIS (Grooper)	G2	Slashdot	SourceForge	PeerSpot	SoftwareReviews	WebinarCare	HubSpot	Gartner	DQ Experts
61	iWay	n	n	n	n	n	n	n	n	n	**Y**	n	n	n	n	n	n	**Y**	n
62	IBM InfoSphere Information Analyzer	**Y**	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n
63	IBM InfoSphere Information Server for Data Quality	n	n	n	n	n	n	n	**Y**	n	n	n	n	**Y**	n	n	n	**Y**	n
64	IBM InfoSphere QualityStage	**Y**	n	**Y**	**Y**	n	n	n	n	n	n	n	n	n	n	n	n	**Y**	n
65	IBM Watson Knowledge Catalog	n	**Y**	**Y**	n	n	n	n	**Y**	**Y**	n	n	n	n	n	n	n	**Y**	n
66	Informatica Address Verification	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	**Y**	n
67	Informatica Axon	n	n	n	n	n	n	n	n	n	n	n	n	**Y**	n	n	n	n	n
68	Informatica Cloud Data Quality	**Y**	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n
69	Informatica Data as a Service	**Y**	**Y**	n	**Y**	n	**Y**	n	**Y**	**Y**	n	n	n	**Y**	**Y**	**Y**	n	**Y**	n
70	Informatica Data Engineering Quality	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	**Y**	n
71	Informatica Enterprise Data Catalog	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	**Y**	n
72	Informatica Master Data Management	**Y**	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n
73	ClearCore	n	**Y**	n	n	**Y**	**Y**	n	n	n	n	n	n	n	n	n	n	n	n
74	OpenDQ	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	**Y**	n

Y - *"yes"*, n - *"no"*

Continuation of Appendix IIa

ID	Tool	Academic papers	Datamation	Simplilearn	TechTarget	Solutions Review	TechRepublic	Geekflare	TrustRadius	BIS (Grooper)	G2	Slashdot	SourceForge	PeerSpot	SoftwareReviews	WebinarCare	HubSpot	Gartner	DQ Experts
75	Enlighten	n	n	n	n	n	n	n	n	n	n	**Y**	**Y**	n	n	n	n	n	n
76	FinScan	n	n	n	n	n	n	n	n	**Y**	n	n	n	n	n	n	n	**Y**	n
77	Synchronos	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	**Y**	n
78	Insycle	n	n	n	n	**Y**	n	n	n	n	n	n	n	n	n	n	n	n	n
79	IQ Office	n	n	n	n	n	n	**Y**	n	n	**Y**	n	n	n	n	n	**Y**	n	n
80	Introhive	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	**Y**	n
81	Irion EDM	n	n	n	n	n	n	n	n	n	**Y**	n	n	n	n	**Y**	n	n	n
82	Data Quality Solution	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	**Y**	n
83	Deduplix	**Y**	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n
84	Scrubbix	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	**Y**	n
85	LiTech Data Quality Management	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	**Y**	n
86	Loqate	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	**Y**
87	Melissa Data Data Profiler	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	**Y**	n
88	Melissa Data Data Quality	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	**Y**	n
89	Melissa Data Data Quality Components for SSIS	n	n	n	n	n	n	n	n	n	n	n	n	**Y**	n	n	n	n	n
90	Melissa Data Global Data Quality Suite	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	**Y**	n
91	Melissa Data MatchUp	n	n	n	n	n	n	n	**Y**	n	n	n	n	n	n	**Y**	n	**Y**	n
92	Melissa Data Personator	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	**Y**	n

Y - *"yes"*, n - *"no"*

Continuation of Appendix IIa																			
ID	Tool	Academic papers	Datamation	Simplilearn	TechTarget	Solutions Review	TechRepublic	Geekflare	TrustRadius	BIS (Grooper)	G2	Slashdot	SourceForge	PeerSpot	SoftwareReviews	WebinarCare	HubSpot	Gartner	DQ Experts
93	Melissa Data Web APIs by Melissa	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	**Y**	n
94	Microsoft Data Quality	n	n	n	n	n	n	n	n	n	n	**Y**	**Y**	n	n	n	n	n	n
95	MIOvantage	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	**Y**	n
96	Monte Carlo Data Observability Platform	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	**Y**	n
97	Nintex	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	**Y**	n
98	Datamartist	n	n	n	n	n	n	n	n	n	n	**Y**	**Y**	n	n	n	n	n	n
99	RevOps Data Automation Cloud	**Y**	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n
100	Oracle Cloud Infrastructure Data Catalog	n	n	n	n	n	n	n	n	n	**Y**	n	n	n	n	n	n	n	n
101	Oracle Enterprise Data Quality	n	n	**Y**	n	**Y**	n	n	n	n	n	n	n	n	n	n	n	n	n
102	OvalEdge	**Y**	n	**Y**	n	n	n	n	**Y**	**Y**	n	n	n	**Y**	n	n	n	**Y**	n
103	Rapid Data Profiling	n	n	n	n	n	n	n	**Y**	n	n	n	n	n	n	n	n	n	n
104	Self-Service Data Preparation	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	**Y**	n
105	Intelligent Data Quality Management	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	**Y**	n
106	Spectrum Technology Platform	n	n	n	n	n	n	n	n	n	n	n	n	n	**Y**	n	n	**Y**	n
107	Duplicate Check for Salesforce	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	**Y**	n

Y - *"yes"*, n - *"no"*

Continuation of Appendix IIa

ID	Tool	Academic papers	Datamation	Simplilearn	TechTarget	Solutions Review	TechRepublic	Geekflare	TrustRadius	BIS (Grooper)	G2	Slashdot	SourceForge	PeerSpot	SoftwareReviews	WebinarCare	HubSpot	Gartner	DQ Experts
108	PostGrid Address Verification	n	n	n	n	n	n	n	n	n	**Y**	n	n	n	n	n	n	n	n
109	Precisely Data360	n	n	n	n	n	n	n	**Y**	n	n	n	n	n	n	n	n	n	n
110	Precisely Spectrum Quality	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	**Y**	n
111	Precisely Trillium Quality	n	n	n	n	n	n	n	**Y**	n	n	n	n	n	n	n	n	**Y**	n
112	Entity Resolution and Data Intelligence Tools	**Y**	**Y**	n	**Y**	**Y**	**Y**	n	n	n	n	n	n	n	n	n	n	**Y**	n
113	rgOne	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	**Y**	n
114	DataLever	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	**Y**	n
115	SAP Address and Geocoding Directories	**Y**	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n
116	SAP Data Intelligence	n	n	n	n	n	n	n	**Y**	n	n	n	n	n	n	n	n	n	n
117	SAP Data Quality Management, microservices for location data	n	n	n	**Y**	n	n	n	n	n	n	n	n	n	n	n	n	n	n
118	SAP Data Services	n	n	n	n	n	n	n	**Y**	n	n	n	n	**Y**	n	**Y**	n	**Y**	n
119	SAP Information Steward	n	n	n	n	n	n	n	**Y**	**Y**	n	n	n	**Y**	**Y**	n	n	**Y**	n
120	SAP Master Data Governance	**Y**	n	n	n	n	n	n	n	n	n	n	n	**Y**	n	n	n	**Y**	n
121	SAS Data Loader for Hadoop	n	n	n	n	**Y**	n	n	n	n	n	n	n	n	n	n	n	n	n
122	SAS Data Management	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	**Y**	n

Y - *"yes"*, n - *"no"*

Continuation of Appendix IIa

ID	Tool	Academic papers	Datamation	Simplilearn	TechTarget	Solutions Review	TechRepublic	Geekflare	TrustRadius	BIS (Grooper)	G2	Slashdot	SourceForge	PeerSpot	SoftwareReviews	WebinarCare	HubSpot	Gartner	DQ Experts
123	SAS Data Quality	n	**Y**	**Y**	n	n	n	n	n	**Y**	n	n	n	**Y**	n	n	**Y**	**Y**	n
124	SAS Data Quality Accelerator for Teradata	**Y**	n	**Y**	**Y**	n	n	n	n	**Y**	n	n	n	n	**Y**	n	**Y**	**Y**	n
125	SAS Dataflux	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	**Y**	n
126	Semarchy xDM	**Y**	n	n	n	n	n	n	**Y**	n	n	n	n	n	n	n	n	n	n
127	Pentaho Kettle	n	n	n	n	n	n	n	n	n	n	**Y**	**Y**	n	n	n	n	n	n
128	Masterpiece -> SpheraCloud Platform	**Y**	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n
129	SQL Power Architect	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	**Y**	n
130	SQL Power DQguru	**Y**	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n
131	Stratio Augmented Data Fabric Platform	**Y**	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n
132	Syncari	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	**Y**	n
133	Syniti Knowledge Platform	n	n	n	n	n	n	n	n	n	**Y**	n	n	n	n	**Y**	n	n	n
134	Syniti Master Data Management	n	n	n	n	n	n	n	n	**Y**	n	n	n	n	n	n	n	**Y**	n
135	Syniti Match	n	n	**Y**	n	**Y**	n	n	n	n	n	n	n	n	n	n	n	n	n
136	RingLead Platform	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	**Y**	n
137	ZoomInfo OperationsOS	n	n	n	n	n	n	n	**Y**	n	n	n	n	n	n	**Y**	n	**Y**	n
138	Tale Of Data	n	n	n	n	n	n	**Y**	n	n	n	n	n	n	**Y**	n	n	n	n

Y - *"yes"*, n - *"no"*

Continuation of Appendix IIa

ID	Tool	Academic papers	Datamation	Simplilearn	TechTarget	Solutions Review	TechRepublic	Geekflare	TrustRadius	BIS (Grooper)	G2	Slashdot	SourceForge	PeerSpot	SoftwareReviews	WebinarCare	HubSpot	Gartner	DQ Experts
139	Talend Data Fabric	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	**Y**	n
140	Talend Data Preparation	n	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	n	n	n	n	n	n	n	n	n	**Y**	n
141	Talend Data Stewardship	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	**Y**	n
142	Talend Open Studio for Data Quality	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	**Y**	n
143	Talend Platform for Data Management (Legacy)	**Y**	n	n	n	n	n	n	**Y**	**Y**	n	n	n	**Y**	**Y**	n	**Y**	**Y**	n
144	TIBCO Clarity	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	**Y**	n
145	TIBCO (Cloud) EBX	n	**Y**	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n
146	iCEDQ	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	**Y**	n
147	Alteryx Designer Cloud	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	**Y**	n
148	Uniserv	n	n	n	n	n	n	n	n	n	n	n	n	n	n	**Y**	n	n	n
149	DataFuse	**Y**	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n
150	DemandTools	n	n	n	n	n	n	n	**Y**	n	n	n	n	n	n	n	n	n	n
151	Clean & Match Enterprise	n	**Y**	n	n	n	n	n	**Y**	n	**Y**	n	n	n	n	**Y**	n	n	n

Y - *"yes"*, n - *"no"*

IIb. DQ Tools: Publications as Sources

	Authors	Title	Year	Source title	DQ Tools	Citation
K1	Ehrlinger L., Wöß W.	A Survey of Data Quality Measurement and Monitoring Tools	2022	Frontiers in Big Data	Aggregate Profiler Apache Griffin Ataccama ONE DataCleaner by Human Inference Datamartist by nModal Solutions Inc. Experian Pandora Informatica Data Quality IBM InfoSphere Information Server for Data Quality InfoZoom by humanIT Software GmbH MobyDQ OpenRefine and MetricDoc Oracle Enterprise Data Quality Talen Open Studio for Data Quality SAS Data Quality SAP Information Steward Data Quality Solution by ISO Professional Services dspCompose by BackOffice Associates GmbH	[EW22]

K1 - *keyword combination 1*: *"data quality tool" OR "data quality software"*

K2 - *keyword combination 2*: *("information quality" OR "data quality") AND ("software" OR "tool" OR "application") AND "data quality rule"*

Continuation of Appendix IIb

	Authors	Title	Year	Source title	DQ Tools	Citation
K1	Ehrlinger L., Gindlhumer A., Huber L.-M., Wöß W.	DQ-MeeRKat: Automating data quality monitoring with a reference-data-profile-annotated knowledge graph	2021	Proceedings of the 10th International Conference on Data Science, Technology and Applications, DATA 2021	Oracle EDQ SAS Talend Informatica DQ-MeeRKat Holodetect	[EGHW21]
K1	Georgieva P., Nikolova E., Orozova D.	Data cleaning techniques in detecting tendencies in software engineering	2020	2020 43rd International Convention on Information, Communication and Electronic Technology, MIPRO 2020 - Proceedings	-	
K1	Azeroual O., Lewoniewski W.	How to inspect and measure data quality about scientific publications: Use case of Wikipedia and CRIS databases	2020	Algorithms	DataCleaner (https://datacleaner.org/)	[AL20]
K1	Masó J., Julia N., Zabala A., Prat E., Kwast J.V.D., Domingo-Marimon C.	Assess citizen science based land cover maps with remote sensing products: The Ground Truth 2.0 data quality tool	2020	Proceedings of SPIE - The International Society for Optical Engineering	-	
K1	Houston L., Probst Y., Yu P., Martin A.	Exploring data quality management within clinical trials	2018	Applied Clinical Informatics	-	

K1 - *keyword combination 1*: *"data quality tool" OR "data quality software"*

K2 - *keyword combination 2*: *("information quality" OR "data quality") AND ("software" OR "tool" OR "application") AND "data quality rule"*

Continuation of Appendix IIb						
	Authors	**Title**	**Year**	**Source title**	**DQ Tools**	**Citation**
K1	Brennan R.	Challenges for value-driven semantic data quality management	2017	ICEIS 2017 - Proceedings of the 19th International Conference on Enterprise Information Systems	-	
K1	Božić B., Brennan R., Feeney K.C., Mendel-Gleason G.	Describing reasoning results with RVO, the reasoning violations ontology	2016	CEUR Workshop Proceedings	-	
K1	Venkatesh Pulla V.S., Varo C., Al M.	Open source data quality tools: Revisited	2016	Advances in Intelligent Systems and Computing	Talend Open Studio DataCleaner WinPure Data Preparator Data Match DataMartist Pentaho Kettle SQL Power Architect SQL Power DQguru DQAnalyzer	[PVA16]
K1	Abdullah N., Ismail S.A., Sophiayati S., Sam S.M.	Data quality in big data: A review	2015	International Journal of Advances in Soft Computing and its Applications	-	

K1 - *keyword combination 1*: *"data quality tool" OR "data quality software"*

K2 - *keyword combination 2*: *("information quality" OR "data quality") AND ("software" OR "tool" OR "application") AND "data quality rule"*

Continuation of Appendix IIb

	Authors	Title	Year	Source title	DQ Tools	Citation
K1	Woodall P., Oberhofer M., Borek A.	A classification of data quality assessment and improvement methods	2014	International Journal of Information Quality	SAS dataflux Informatica Trillium software SAP IBM Pitney Bowes Software Oracle Datactics DataMentors RedPoint-DataLever Uniserv Innovative Systems Human Inference Talend Information Builders/iWay Ataccama	[WOB14a]
K1	Chiang F., Wang Y.	Repairing integrity rules for improved data quality	2014	International Journal of Information Quality	-	
K1	Guerra-García C., Caballero I., Piattini M.	Capturing data quality requirements for web applications by means of DQ-WebRE	2013	Information Systems Frontiers	-	
K1	Capirossi J., Rabier P.	An enterprise architecture and data quality framework	2013	Advances in Intelligent Systems and Computing	-	

K1 - *keyword combination 1*: *"data quality tool" OR "data quality software"*

K2 - *keyword combination 2*: *("information quality" OR "data quality") AND ("software" OR "tool" OR "application") AND "data quality rule"*

Continuation of Appendix IIb

	Authors	Title	Year	Source title	DQ Tools	Citation
K2	Song S., Gao F., Huang R., Wang C.	Data Dependencies Extended for Variety and Veracity: A Family Tree	2022	IEEE Transactions on Knowledge and Data Engineering	-	
K2	Bronselaer A., Boeckling T., Pattyn F.	Dynamic repair of categorical data with edit rules	2022	Expert Systems with Applications	-	
K2	Nulhusna R., Taufiq N.F., Ruldeviyani Y.	Strategy to Improve Data Quality Management: A Case Study of Master Data at Government Organization in Indonesia	2022	Proceeding - 2022 International Symposium on Information Technology and Digital Innovation: Technology Innovation During Pandemic, ISITDI 2022	-	
K2	Hasan F.F., Hameed S.J.	The Perspective of Data Quality Rules in Google Forms	2022	ISMSIT 2022 - 6th International Symposium on Multidisciplinary Studies and Innovative Technologies, Proceedings	-	
K2	Chaudhary K., D'Spain K., Khanal S., Nguyen K., Pham L., Lall G., Trieu T., Kadiyala K., Wei B.	Toyota Financial Services Data Portal	2022	IST 2022 - IEEE International Conference on Imaging Systems and Techniques, Proceedings	Informatica Axon Informatica DQ Informatica EDC	[CDK+22]

K1 - *keyword combination 1*: *"data quality tool" OR "data quality software"*

K2 - *keyword combination 2*: *("information quality" OR "data quality") AND ("software" OR "tool" OR "application") AND "data quality rule"*

Continuation of Appendix IIb						
	Authors	**Title**	**Year**	**Source title**	**DQ Tools**	**Citation**
K2	Valencia-Parra Á., Parody L., Varela-Vaca Á.J., Caballero I., Gómez-López M.T.	DMN4DQ: When data quality meets DMN	2021	Decision Support Systems	-	
K2	Lettner C., Stumptner R., Fragner W., Rauchenzauner F., Ehrlinger L.	DaQL 2.0: Measure Data Quality based on Entity Models	2021	Procedia Computer Science	-	
K2	Loetpipatwanich S., Vichitthamaros P.	Sakdas: A Python Package for Data Profiling and Data Quality Auditing	2020	2020 1st International Conference on Big Data Analytics and Practices, IBDAP 2020	-	
K2	Heine F., Kleiner C., Oelsner T.	A dsl for automated data quality monitoring	2020	Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)	-	

K1 - *keyword combination 1*: *"data quality tool" OR "data quality software"*

K2 - *keyword combination 2*: *("information quality" OR "data quality") AND ("software" OR "tool" OR "application") AND "data quality rule"*

Continuation of Appendix IIb						
	Authors	**Title**	**Year**	**Source title**	**DQ Tools**	**Citation**
K2	Boeckling T., Bronselaer A., de Tré G.	Mining data quality rules based on T-dependence	2020	Proceedings of the 11th Conference of the European Society for Fuzzy Logic and Technology, EUSFLAT 2019	-	
K2	Liu H., Wang X., Lei S., Zhang X., Liu W., Qin M.	A rule based data quality assessment architecture and application for electrical data	2019	ACM International Conference Proceeding Series	-	
K2	van Cruchten R.M.E.	Data quality in process mining: A rule-based Approach	2019	CEUR Workshop Proceedings	-	
K2	Sun J., Li J.	Discovery of MicroDependencies	2019	IEEE Access	-	
K2	Juddoo S.	Overview of data quality challenges in the context of Big Data	2016	2015 International Conference on Computing, Communication and Security, ICCCS 2015	-	
K2	Hoel E., Bakalov P., Kim S., Brown T.	Moving beyond transportation: Utility network management	2015	GIS: Proceedings of the ACM International Symposium on Advances in Geographic Information Systems	-	
K2	Abdo A.S., Salem R.K., Abdul-Kader H.M.	Efficient Dependable Rules Generation Approach for Data Quality Enhancement	2015	25th International Conference on Computer Theory and Applications, ICCTA 2015 - Proceedings	-	

K1 - *keyword combination 1*: *"data quality tool" OR "data quality software"*

K2 - *keyword combination 2*: *("information quality" OR "data quality") AND ("software" OR "tool" OR "application") AND "data quality rule"*

Continuation of Appendix IIb						
	Authors	**Title**	**Year**	**Source title**	**DQ Tools**	**Citation**
K2	Chu X., Ilyas I.F., Papotti P., Ye Y.	RuleMiner: Data quality rules discovery	2014	Proceedings - International Conference on Data Engineering	-	
K2	Zanzi A., Trombetta A.	Data quality evaluation of scientific datasets a case study in a policy support context	2013	DATA 2013 - Proceedings of the 2nd International Conference on Data Technologies and Applications	-	
K2	Valeva V., Roach P.A.	A proposed methodology for automated product data quality rule generation	2013	28th International Conference on Computers and Their Applications 2013, CATA 2013	-	
K2	Dallachiesat M., Ebaid A., Eldawy A., Elmagarmid A., Ilyas I.F., Ouzzani M., Tang N.	NADEEF: A commodity data cleaning system	2013	Proceedings of the ACM SIGMOD International Conference on Management of Data	-	
K2	Gao J., Woodall P., Koronios A., Parlikad A.K.	Data profiling challenges in engineering asset management data – Conceptual design for next generation data profiling software	2013	Proceedings of the 18th International Conference on Information Quality, ICIQ 2013	-	

K1 - *keyword combination 1*: *"data quality tool" OR "data quality software"*

K2 - *keyword combination 2*: *("information quality" OR "data quality") AND ("software" OR "tool" OR "application") AND "data quality rule"*

III. DQ Tools: Source of the Information

ID	Tool	Official Website	Video	Additional Information	Trialability	Documentation	Level of Information	Selection 1
1	Data Preparator				Not interested	-	-	n
2	Holodetect	Link		Link	Open source	**Y**	Good	**Y**
3	MetricDoc	Link			Open source	**Y**	Good	**Y**
4	DataMentors				Not interested	-	-	n
5	DQ-MeeRKat	Link		Link	Open source	**Y**	Good	**Y**
6	MobyDQ	Link			Open source	**Y**	Good	**Y**
7	Great Expectations	Link		Link	Open source	**Y**	Good	**Y**
8	AbInitio Enterprise Data Platform	Link			Not available	n	Good	**Y**
9	Acceldata	Link	Link	Link	Request a free trial	**Y**	Good	**Y**
10	DQ*Plus Enterprise Suite	Link			Not available	n	Partial	**Y**
11	Amperity CDP	Link	Link		Not interested	-	-	n
12	Anomalo	Link	Link		Demo available	n	Good	**Y**
13	Apache Griffin	Link		Link	Open source	**Y**	Good	**Y**
14	Aggregate Profiler	Link			Open source	**Y**	Good	**Y**
15	Attaccama DQ-Analyzer	Link			Open source	**Y**	Good	**Y**
16	Ataccama ONE	Link	Link	Link	Demo available	n	Good	**Y**
17	dspCompose	Link			Not available	**Y**	Good	**Y**
18	CRM Cleaning	Link		Link	Request a demo	n	Good	**Y**

Y - *"yes"*, n - *"no"*

Selection 1: Y - *"included"*, n - *"excluded"*

Continuation of Appendix III

ID	Tool	Official Website	Video	Additional Information	Trialability	Documentation	Level of Information	Selection 1
19	Black Tiger Platform	Link			Not available	n	Low	n
20	ChainSys dataZen	Link	Link		Not interested	-	-	n
21	Smart Data Platform	Link	Link	Link	Request a free trial	n	Good	**Y**
22	Claravine	Link			Not interested	-	-	n
23	ClearAnalytics	Link		Link	Not interested	-	-	n
24	Cloudingo	Link			Free trial available	n	Good	**Y**
25	Collibra Platform	Link	Link	Link	Demo available	**Y**	Good	**Y**
26	Cribl Stream	Link			Not interested	-	-	n
27	CuriumDQM	Link			Not available	n	Partial	**Y**
28	D&B Connect	Link			Request a demo	n	Partial	**Y**
29	D&B Optimizer	Link			Request a demo	n	Partial	**Y**
30	DataMatch Enterprise	Link			Free trial available	**Y**	Good	**Y**
31	Datactics Self-Service Data Quality Platform	Link		Link	Request a demo	n	Partial	**Y**
32	Dataedo	Link			Request a free trial	**Y**	Good	**Y**
33	DataStreams Platform	Link			Request a demo	n	Low	n
34	Ultimate Data Export	Link			Not interested	-	-	n
35	Datiris Profiler	Link			Not available	-	-	n

Y - *"yes"*, n - *"no"*

Selection 1: Y - *"included"*, n - *"excluded"*

Continuation of Appendix III

ID	Tool	Official Website	Video	Additional Information	Trialability	Documentation	Level of Information	Selection 1
36	Dedupely	Link			Free trial available	n	Good	**Y**
37	MyDataQ	Link			Request a demo	n	Partial	**Y**
38	DQE One	Link			Request a demo	n	Partial	**Y**
39	DQLABS Platform	Link		Link	Request a demo	n	Good	**Y**
40	Duco Platform	Link			Request a demo	n	Good	**Y**
41	DvSum	Link		Link	Request a demo	n	Good	**Y**
42	Edge Delta	Link			Not interested	-	-	n
43	Exmon	Link			Request a demo	n	Good	**Y**
44	Experian Aperture Data Studio	Link	Link	Link	Request a free trial	**Y**	Good	**Y**
45	Experian DataArc360	Link			Not available	n	Good	**Y**
46	Experian Email Validation	Link			Not available	n	Good	**Y**
47	Experian Namesearch	Link		Link	Not available	n	Good	**Y**
48	Experian Pandora (Legacy)	Link			Not available	-	-	n
49	Experian Phone Validation	Link			Not available	n	Good	**Y**
50	Experian Prospect IQ	Link			Not available	**Y**	Good	**Y**
51	Flatfile	Link		Link	Request a demo	**Y**	Good	**Y**
52	Global IDs Data Quality Suites	Link			Request a demo	n	Good	**Y**

Y - *"yes"*, n - *"no"*

Selection 1: Y - *"included"*, n - *"excluded"*

Continuation of Appendix III								
ID	Tool	Official Website	Video	Additional Information	Trialability	Documentation	Level of Information	Selection 1
53	OpenRefine	Link			Open source	**Y**	Good	**Y**
54	matchIT Data Quality Solutions				Not interested	-	-	n
55	HERE Platform	Link			Not interested	-	-	n
56	HubSpot Operations Hub	Link			Not interested	-	-	n
57	DataCleaner	Link		Link	Open source	**Y**	Good	**Y**
58	InfoZoom	Link	Link		Request a free trial	**Y**	Partial	**Y**
59	ibi Data Quality	Link	Link	Link	Not available	n	Good	**Y**
60	ibi Omni-Gen	Link			Not interested	-	-	n
61	iWay	Link			Not interested	-	-	n
62	IBM InfoSphere Information Analyzer	Link			Not interested	-	-	n
63	IBM InfoSphere Information Server for Data Quality	Link		Link	Not available	**Y**	Good	**Y**
64	IBM InfoSphere QualityStage	Link		Link	Not interested	-	-	n
65	IBM Watson Knowledge Catalog	Link		Link	Not interested	-	-	n
66	Informatica Address Verification	Link			Not available	**Y**	Good	**Y**
67	Informatica Axon	Link	Link		Not interested	-	-	n

Y - *"yes"*, n - *"no"*

Selection 1: Y - *"included"*, n - *"excluded"*

Continuation of Appendix III

ID	Tool	Official Website	Video	Additional Information	Trialability	Documentation	Level of Information	Selection 1
68	Informatica Cloud Data Quality	Link	Link	Link	Not available	n	Good	**Y**
69	Informatica Data as a Service	Link			Not available	**Y**	Good	**Y**
70	Informatica Data Engineering Quality	Link		Link	Not available	**Y**	Good	**Y**
71	Informatica Enterprise Data Catalog	Link			Not interested	-	-	n
72	Informatica Master Data Management	Link	Link	Link	Demo available	n	Good	**Y**
73	ClearCore	Link		Link	Request a demo	n	Partial	**Y**
74	OpenDQ	Link			Request a demo	n	Low	n
75	Enlighten	Link		Link	Request a demo	n	Partial	**Y**
76	FinScan	Link			Not interested	-	-	n
77	Synchronos				Not interested	-	-	n
78	Insycle	Link			Free trial available	n	Good	**Y**
79	IQ Office	Link			Request a free trial	n	Partial	**Y**
80	Introhive	Link			Not interested	-	-	n
81	Irion EDM	Link	Link		Not interested	-	-	n
82	Data Quality Solution	Link			Request a demo	n	Partial	**Y**
83	Deduplix	Link			Request a demo	n	Low	n

Y - *"yes"*, n - *"no"*

Selection 1: Y - *"included"*, n - *"excluded"*

Continuation of Appendix III

ID	Tool	Official Website	Video	Additional Information	Trialability	Documentation	Level of Information	Selection 1
84	Scrubbix	Link			Request a demo	n	Partial	**Y**
85	LiTech Data Quality Management	Link		Link	Request a free trial	**Y**	Good	**Y**
86	Loqate	Link			Free trial available	**Y**	Good	**Y**
87	Melissa Data Data Profiler	Link			Not interested	-	-	n
88	Melissa Data Data Quality	Link			Not available	-	-	n
89	Melissa Data Data Quality Components for SSIS	Link			Request a demo	**Y**	Good	**Y**
90	Melissa Data Global Data Quality Suite	Link			Not available	-	-	n
91	Melissa Data MatchUp	Link			Not interested	-	-	n
92	Melissa Data Personator	Link			Not interested	-	-	n
93	Melissa Data Web APIs by Melissa	Link		Link	Request a demo	**Y**	Good	**Y**
94	Microsoft Data Quality	Link		Link	Free trial available	**Y**	Good	**Y**
95	MIOvantage	Link		Link	Request a demo	**Y**	Good	**Y**

Y - *"yes"*, n - *"no"*

Selection 1: Y - *"included"*, n - *"excluded"*

Continuation of Appendix III								
ID	Tool	Official Website	Video	Additional Information	Trialability	Documentation	Level of Information	Selection 1
96	Monte Carlo Data Observability Platform	Link		Link	Request a demo	**Y**	Good	**Y**
97	Nintex	Link			Not interested	-	-	n
98	Datamartist	Link			Free trial available	**Y**	Good	**Y**
99	RevOps Data Automation Cloud	Link			Not interested	-	-	n
100	Oracle Cloud Infrastructure Data Catalog	Link	Link		Not interested	-	-	n
101	Oracle Enterprise Data Quality	Link			Free trial available	**Y**	Partial	**Y**
102	OvalEdge	Link			Request a demo	n	Good	**Y**
103	Rapid Data Profiling	Link			Request a demo	**Y**	Partial	**Y**
104	Self-Service Data Preparation	Link			Not interested	-	-	n
105	Intelligent Data Quality Management	Link			Not available	n	Partial	**Y**
106	Spectrum Technology Platform	Link		Link	Not available	n	Partial	**Y**
107	Duplicate Check for Salesforce	Link			Free trial available	n	Good	**Y**
108	PostGrid Address Verification	Link			Request a demo	**Y**	Good	**Y**
109	Precisely Data360	Link			Request a demo	**Y**	Good	**Y**

Y - *"yes"*, n - *"no"*

Selection 1: Y - *"included"*, n - *"excluded"*

Continuation of Appendix III

ID	Tool	Official Website	Video	Additional Information	Trialability	Documentation	Level of Information	Selection 1
110	Precisely Spectrum Quality	Link			Request a demo	n	Partial	**Y**
111	Precisely Trillium Quality	Link		Link	Request a demo	**Y**	Good	**Y**
112	Entity Resolution and Data Intelligence Tools	Link			Not interested	-	-	n
113	rgOne	Link	Link		Request a demo	n	Partial	**Y**
114	DataLever				Not available	-	-	n
115	SAP Address and Geocoding Directories	Link		Link	Not available	**Y**	Good	**Y**
116	SAP Data Intelligence	Link			Not interested	-	-	n
117	SAP Data Quality Management, microservices for location data	Link		Link	Request a demo	**Y**	Good	**Y**
118	SAP Data Services	Link		Link	Request a demo	**Y**	Partial	**Y**
119	SAP Information Steward	Link	Link	Link	Not available	**Y**	Good	**Y**
120	SAP Master Data Governance	Link	Link	Link	Demo available	**Y**	Partial	**Y**
121	SAS Data Loader for Hadoop	Link		Link	Request a demo	**Y**	Good	**Y**
122	SAS Data Management	Link		Link	Request a demo	**Y**	Good	**Y**
123	SAS Data Quality	Link		Link	Request a demo	**Y**	Good	**Y**

Y - *"yes"*, n - *"no"*

Selection 1: Y - *"included"*, n - *"excluded"*

Continuation of Appendix III								
ID	Tool	Official Website	Video	Additional Information	Trialability	Documentation	Level of Information	Selection 1
124	SAS Data Quality Accelerator for Teradata	Link	Link	Link	Not available	**Y**	Good	**Y**
125	SAS Dataflux	Link		Link	Request a demo	**Y**	Good	**Y**
126	Semarchy xDM	Link			Request a demo	n	Good	**Y**
127	Pentaho Kettle	Link			Not interested	-	-	n
128	Masterpiece -> SpheraCloud Platform	Link			Not interested	-	-	n
129	SQL Power Architect	Link			Not interested	-	-	n
130	SQL Power DQguru	Link			Open source	**Y**	Good	**Y**
131	Stratio Augmented Data Fabric Platform	Link			Request a demo	n	Partial	**Y**
132	Syncari	Link			Not interested	-	-	n
133	Syniti Knowledge Platform	Link	Link	Link	Request a demo	**Y**	Good	**Y**
134	Syniti Master Data Management	Link		Link	Not interested	-	-	n
135	Syniti Match	Link		Link	Not interested	-	-	n
136	RingLead Platform	Link			Open source	**Y**	Good	**Y**
137	ZoomInfo OperationsOS	Link			Request a demo	n	Partial	**Y**

Y - *"yes"*, n - *"no"*

Selection 1: Y - *"included"*, n - *"excluded"*

Continuation of Appendix III

ID	Tool	Official Website	Video	Additional Information	Trialability	Documentation	Level of Information	Selection 1
138	Tale Of Data	Link			Request a free trial	n	Good	**Y**
139	Talend Data Fabric	Link	Link		Request a free trial	n	Partial	**Y**
140	Talend Data Preparation	Link			Request a free trial	n	Partial	**Y**
141	Talend Data Stewardship	Link		Link	Request a demo	**Y**	Partial	**Y**
142	Talend Open Studio for Data Quality	Link			Open source	**Y**	Low	n
143	Talend Platform for Data Management (Legacy)	Link			Not available	-	-	n
144	TIBCO Clarity	Link		Link	Request a demo	**Y**	Good	**Y**
145	TIBCO (Cloud) EBX	Link	Link		Not interested	-	-	n
146	iCEDQ	Link	Link		Request a demo	n	Good	**Y**
147	Alteryx Designer Cloud	Link			Not interested	-	-	n
148	Uniserv	Link			Request a free trial	n	Good	**Y**
149	DataFuse	Link			Not available	n	Partial	**Y**
150	DemandTools	Link			Request a free trial	n	Partial	**Y**
151	Clean & Match Enterprise	Link			Free trial available	n	Good	**Y**

Y - *"yes"*, n - *"no"*

Selection 1: Y - *"included"*, n - *"excluded"*

IV. DQ Tools: DQ and Other Features

ID	Tool	Data Profiling	Custom DQ Rules	DQ Rule Definition in SQL	DQ Dimensions Used	DQ Rules Repository	Erroneous Records Shown	DQ Report Creation	DQ Dashboard	Data Match Detection	Anomaly Detection	DQ Rule Detection	Data Cleansing	Data Enrichment	Master Data Management	Data Lineage	Data Catalogue	Data Semantics discovery	Data Integration	Selection 2
2	Holodetect	**Y**	n	n	n	n	n	n	n	n	**Y**	n	**Y**	**Y**	n	n	n	n	n	n
3	MetricDoc	n	**Y**	n	**Y**	n	**Y**	n	n	n	n	n	n	n	n	n	n	n	n	n
5	DQ-MeeRKat	**Y**	n	n	n	**Y**	n	n	**Y**	n	**Y**	n	n	n	n	n	n	n	n	n
6	MobyDQ	n	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	n	**Y**	n	n	n	n	n	n	n	n	Y*
7	Great Expectations	**Y**	**Y**	n	n	**Y**	**Y**	**Y**	n	n	n	n	n	n	n	n	n	n	n	n
8	AbInitio Enterprise Data Platform	**Y**	**Y**	n	**Y**	**Y**	**Y**	**Y**	n	**Y**	n	**Y**	**Y**	n	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**
9	Acceldata	**Y**	n	n	n	n	n	n	**Y**	n	**Y**	n	n	n	n	**Y**	n	n	**Y**	n
10	DQ*Plus Enterprise Suite	n	n	n	n	n	n	n	n	**Y**	n	n	**Y**	**Y**	n	n	n	n	n	n
12	Anomalo	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	n	**Y**	n	**Y**	n	n	n	n	n	n	n	n	Y*
13	Apache Griffin	**Y**	**Y**	n	**Y**	**Y**	n	**Y**	n	n	n	n	n	n	n	n	n	n	n	n
14	Aggregate Profiler	**Y**	**Y**	n	n	n	**Y**	**Y**	n	**Y**	n	n	**Y**	**Y**	n	n	**Y**	n	n	n
15	Attaccama DQAnalyzer	**Y**	**Y**	n	n	**Y**	**Y**	n	n	**Y**	n	n	**Y**	n	n	n	n	n	n	n

Y - *"yes"*, n - *"no"*

Selection 2: Y - *"included"*, Y* - *"included as alternative"*, n - *"excluded"*

Continuation of Appendix IV

ID	Tool	Data Profiling	Custom DQ Rules	DQ Rule Definition in SQL	DQ Dimensions Used	DQ Rules Repository	Erroneous Records Shown	DQ Report Creation	DQ Dashboard	Data Match Detection	Anomaly Detection	DQ Rule Detection	Data Cleansing	Data Enrichment	Master Data Management	Data Lineage	Data Catalogue	Data Semantics discovery	Data Integration	Selection 2
16	Ataccama ONE	**Y**	**Y**	n	n	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	n	**Y**
17	dspCompose	n	n	n	n	n	n	n	n	n	n	n	**Y**	**Y**	n	n	n	n	n	n
18	CRM Cleaning	n	n	n	n	n	n	n	n	n	n	n	**Y**	**Y**	n	n	n	n	n	n
21	Smart Data Platform	**Y**	**Y**	n	n	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	n	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	Y*
24	Cloudingo	**Y**	n	n	n	n	n	n	n	**Y**	n	n	**Y**	**Y**	n	n	n	n	n	n
25	Collibra Platform	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	n	n	n	**Y**	**Y**	**Y**	**Y**	**Y**
27	CuriumDQM	**Y**	**Y**	n	n	**Y**	**Y**	**Y**	**Y**	n	n	n	**Y**	n	**Y**	n	n	n	**Y**	n
28	D&B Connect	**Y**	n	n	n	n	n	**Y**	**Y**	**Y**	n	n	**Y**	**Y**	**Y**	n	n	n	**Y**	n
29	D&B Optimizer	**Y**	n	n	n	n	n	n	**Y**	n	n	n	**Y**	**Y**	**Y**	n	n	n	n	n
30	DataMatch Enterprise	**Y**	**Y**	n	n	**Y**	**Y**	n	n	**Y**	n	n	**Y**	**Y**	n	n	n	n	n	n
31	Datactics Self-Service Data Quality Platform	**Y**	**Y**	n	**Y**	**Y**	**Y**	**Y**	**Y**	n	n	n	**Y**	**Y**	**Y**	**Y**	**Y**	n	n	n
32	Dataedo	**Y**	n	n	n	n	n	n	n	n	n	n	n	n	n	**Y**	**Y**	n	n	n

Y - *"yes"*, n - *"no"*

Selection 2: Y - *"included"*, Y* - *"included as alternative"*, n - *"excluded"*

Continuation of Appendix IV

ID	Tool	Data Profiling	Custom DQ Rules	DQ Rule Definition in SQL	DQ Dimensions Used	DQ Rules Repository	Erroneous Records Shown	DQ Report Creation	DQ Dashboard	Data Match Detection	Anomaly Detection	DQ Rule Detection	Data Cleansing	Data Enrichment	Master Data Management	Data Lineage	Data Catalogue	Data Semantics discovery	Data Integration	Selection 2
36	Dedupely	n	n	n	n	n	n	n	n	**Y**	n	n	n	n	**Y**	n	n	n	n	n
37	MyDataQ	n	n	n	n	n	n	n	n	**Y**	n	n	**Y**	**Y**	n	n	n	n	n	n
38	DQE One	n	n	n	n	n	n	n	n	**Y**	n	n	**Y**	n	**Y**	n	n	n	n	n
39	DQLABS Platform	**Y**	**Y**	n	**Y**	**Y**	**Y**	**Y**	**Y**	n	**Y**	**Y**	n	n	**Y**	**Y**	**Y**	**Y**	n	**Y**
40	Duco Platform	n	**Y**	n	n	n	n	n	**Y**	n	n	n	**Y**	**Y**	n	n	n	n	n	n
41	DvSum	**Y**	**Y**	n	**Y**	**Y**	**Y**	**Y**	**Y**	n	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	n	**Y**
43	Exmon	n	**Y**	n	**Y**	**Y**	**Y**	**Y**	**Y**	n	n	n	n	n	**Y**	n	n	n	**Y**	n
44	Experian Aperture Data Studio	**Y**	**Y**	n	n	**Y**	n	n	n	**Y**	**Y**	n	**Y**	**Y**	n	n	**Y**	**Y**	**Y**	Y*
45	Experian DataArc360	**Y**	n	n	**Y**	n	**Y**	**Y**	**Y**	**Y**	**Y**	n	**Y**	**Y**	n	n	n	n	**Y**	n
46	Experian Email Validation	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n
47	Experian Namesearch	n	n	n	n	n	n	n	n	**Y**	n	n	**Y**	**Y**	n	n	n	n	n	n
49	Experian Phone Validation	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n

Y - *"yes"*, n - *"no"*

Selection 2: Y - *"included"*, Y* - *"included as alternative"*, n - *"excluded"*

Continuation of Appendix IV

ID	Tool	Data Profiling	Custom DQ Rules	DQ Rule Definition in SQL	DQ Dimensions Used	DQ Rules Repository	Erroneous Records Shown	DQ Report Creation	DQ Dashboard	Data Match Detection	Anomaly Detection	DQ Rule Detection	Data Cleansing	Data Enrichment	Master Data Management	Data Lineage	Data Catalogue	Data Semantics discovery	Data Integration	Selection 2
50	Experian Prospect IQ	n	n	n	n	n	n	n	n	n	n	n	n	**Y**	n	n	n	n	n	n
51	Flatfile	n	n	n	n	n	n	n	n	n	n	n	**Y**	n	n	n	n	n	n	n
52	Global IDs Data Quality Suites	**Y**	**Y**	n	n	**Y**	**Y**	**Y**	**Y**	n	**Y**	**Y**	n	n	**Y**	**Y**	**Y**	**Y**	n	**Y**
53	OpenRefine	n	n	n	n	n	**Y**	n	n	**Y**	n	n	**Y**	**Y**	n	n	n	n	n	n
57	DataCleaner	**Y**	**Y**	n	n	n	n	n	n	**Y**	n	n	**Y**	**Y**	n	n	n	n	n	n
58	InfoZoom	**Y**	**Y**	n	n	**Y**	**Y**	n	n	n	n	n	**Y**	n	n	n	n	n	n	n
59	ibi Data Quality	**Y**	**Y**	n	n	**Y**	**Y**	n	n	**Y**	n	n	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	n
63	IBM InfoSphere Information Server for Data Quality	**Y**	**Y**	n	n	**Y**	**Y**	**Y**	**Y**	**Y**	n	n	**Y**	**Y**	**Y**	**Y**	**Y**	n	n	n
66	Informatica Address Verification	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n
68	Informatica Cloud Data Quality	**Y**	**Y**	n	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	n	n	n	n	n	n	**Y**

Y - *"yes"*, n - *"no"*

Selection 2: Y - *"included"*, Y* - *"included as alternative"*, n - *"excluded"*

Continuation of Appendix IV

ID	Tool	Data Profiling	Custom DQ Rules	DQ Rule Definition in SQL	DQ Dimensions Used	DQ Rules Repository	Erroneous Records Shown	DQ Report Creation	DQ Dashboard	Data Match Detection	Anomaly Detection	DQ Rule Detection	Data Cleansing	Data Enrichment	Master Data Management	Data Lineage	Data Catalogue	Data Semantics discovery	Data Integration	Selection 2
69	Informatica Data as a Service	n	n	n	n	n	n	n	n	n	n	n	n	**Y**	n	n	n	n	n	n
70	Informatica Data Engineering Quality	**Y**	**Y**	n	**Y**	**Y**	**Y**	**Y**	n	**Y**	**Y**	**Y**	**Y**	**Y**	n	n	n	n	n	**Y**
72	Informatica Master Data Management	**Y**	**Y**	n	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**
73	ClearCore	**Y**	n	n	**Y**	n	**Y**	**Y**	n	**Y**	n	n	**Y**	**Y**	**Y**	n	n	n	n	n
75	Enlighten	**Y**	n	n	n	n	**Y**	n	n	**Y**	n	n	**Y**	**Y**	n	n	n	n	**Y**	n
78	Insycle	**Y**	n	n	n	n	n	n	n	**Y**	n	n	**Y**	n	**Y**	n	n	n	n	n
79	IQ Office	**Y**	n	n	n	n	n	n	n	**Y**	n	n	**Y**	**Y**	n	n	n	n	n	n
82	Data Quality Solution	n	n	n	n	n	n	n	n	n	n	n	n	n	**Y**	n	n	n	**Y**	n
84	Scrubbix	**Y**	n	n	n	n	n	n	n	n	n	n	**Y**	**Y**	n	n	n	n	n	n
85	LiTech Data Quality Management	**Y**	**Y**	**Y**	n	**Y**	**Y**	**Y**	**Y**	n	**Y**	n	n	n	n	n	n	n	n	Y*

Y - *"yes"*, n - *"no"*

Selection 2: Y - *"included"*, Y* - *"included as alternative"*, n - *"excluded"*

Continuation of Appendix IV

ID	Tool	Data Profiling	Custom DQ Rules	DQ Rule Definition in SQL	DQ Dimensions Used	DQ Rules Repository	Erroneous Records Shown	DQ Report Creation	DQ Dashboard	Data Match Detection	Anomaly Detection	DQ Rule Detection	Data Cleansing	Data Enrichment	Master Data Management	Data Lineage	Data Catalogue	Data Semantics discovery	Data Integration	Selection 2
86	Loqate	n	n	n	n	n	n	n	n	n	n	n	**Y**	**Y**	n	n	n	n	**Y**	n
89	Melissa Data Data Quality Components for SSIS	**Y**	n	n	n	n	n	n	n	**Y**	**Y**	n	**Y**	**Y**	n	n	n	n	n	n
93	Melissa Data Web APIs by Melissa	**Y**	n	n	n	n	n	n	n	n	n	n	**Y**	**Y**	n	n	n	n	n	n
94	Microsoft Data Quality	**Y**	**Y**	n	**Y**	**Y**	**Y**	n	n	**Y**	n	n	**Y**	**Y**	n	n	n	n	n	n
95	MIOvantage	**Y**	**Y**	n	n	**Y**	n	**Y**	**Y**	**Y**	n	n	**Y**	**Y**	**Y**	n	**Y**	**Y**	**Y**	n
96	Monte Carlo Data Observability Platform	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	n	n	n	n	**Y**	**Y**	**Y**	n	Y*
98	Datamartist	**Y**	n	n	n	n	n	n	n	**Y**	n	n	**Y**	n	n	n	n	n	n	n
101	Oracle Enterprise Data Quality	**Y**	n	n	n	n	n	n	n	**Y**	n	n	**Y**	**Y**	n	n	n	n	n	n
102	OvalEdge	**Y**	**Y**	**Y**	n	**Y**	**Y**	**Y**	**Y**	n	n	n	n	n	n	**Y**	**Y**	**Y**	n	n
103	Rapid Data Profiling	**Y**	n	n	n	n	**Y**	n	n	n	**Y**	n	**Y**	**Y**	n	n	n	n	n	n

Y - *"yes"*, n - *"no"*

Selection 2: Y - *"included"*, Y* - *"included as alternative"*, n - *"excluded"*

Continuation of Appendix IV

ID	Tool	Data Profiling	Custom DQ Rules	DQ Rule Definition in SQL	DQ Dimensions Used	DQ Rules Repository	Erroneous Records Shown	DQ Report Creation	DQ Dashboard	Data Match Detection	Anomaly Detection	DQ Rule Detection	Data Cleansing	Data Enrichment	Master Data Management	Data Lineage	Data Catalogue	Data Semantics discovery	Data Integration	Selection 2
105	Intelligent Data Quality Management	**Y**	n	n	n	n	**Y**	n	n	**Y**	n	n	**Y**	**Y**	**Y**	n	**Y**	n	n	n
106	Spectrum Technology Platform	**Y**	**Y**	n	n	n	**Y**	n	**Y**	**Y**	n	n	**Y**	**Y**	n	n	n	n	n	n
107	Duplicate Check for Salesforce	n	n	n	n	n	n	n	n	**Y**	n	n	**Y**	n	n	n	n	n	n	n
108	PostGrid Address Verification	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n
109	Precisely Data360	**Y**	**Y**	n	**Y**	**Y**	n	**Y**	**Y**	n	**Y**	n	**Y**	n	n	**Y**	**Y**	**Y**	n	Y*
110	Precisely Spectrum Quality	**Y**	**Y**	n	n	n	**Y**	n	n	**Y**	**Y**	**Y**	**Y**	**Y**	n	n	n	n	n	**Y**
111	Precisely Trillium Quality	**Y**	**Y**	n	n	**Y**	**Y**	n	n	**Y**	n	n	**Y**	**Y**	n	n	n	n	n	n
113	rgOne	**Y**	n	n	n	n	n	n	n	**Y**	n	n	**Y**	**Y**	**Y**	n	n	n	**Y**	n
115	SAP Address and Geocoding Directories	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n

Y - *"yes"*, n - *"no"*

Selection 2: Y - *"included"*, Y* - *"included as alternative"*, n - *"excluded"*

Continuation of Appendix IV

ID	Tool	Data Profiling	Custom DQ Rules	DQ Rule Definition in SQL	DQ Dimensions Used	DQ Rules Repository	Erroneous Records Shown	DQ Report Creation	DQ Dashboard	Data Match Detection	Anomaly Detection	DQ Rule Detection	Data Cleansing	Data Enrichment	Master Data Management	Data Lineage	Data Catalogue	Data Semantics discovery	Data Integration	Selection 2
117	SAP Data Quality Management, microservices for location data	n	n	n	n	n	n	n	n	n	n	n	**Y**	**Y**	n	n	n	n	n	n
118	SAP Data Services	**Y**	n	n	n	n	n	n	**Y**	n	n	n	**Y**	n	n	n	n	n	**Y**	n
119	SAP Information Steward	**Y**	**Y**	n	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	n	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**
120	SAP Master Data Governance	n	**Y**	n	n	n	**Y**	n	**Y**	n	n	n	**Y**	n	**Y**	n	n	n	n	n
121	SAS Data Loader for Hadoop	**Y**	n	n	n	n	n	n	n	**Y**	n	n	**Y**	**Y**	n	n	n	n	**Y**	n
122	SAS Data Management	**Y**	**Y**	n	n	**Y**	**Y**	**Y**	**Y**	**Y**	n	n	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	n
123	SAS Data Quality	**Y**	**Y**	n	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	n	n	**Y**	**Y**	**Y**	**Y**	**Y**	n	**Y**	n
124	SAS Data Quality Accelerator for Teradata	n	n	n	n	n	n	n	n	**Y**	n	n	**Y**	**Y**	n	n	n	n	n	n
125	SAS Dataflux	**Y**	**Y**	n	**Y**	**Y**	**Y**	n	**Y**	**Y**	n	n	**Y**	**Y**	**Y**	n	**Y**	n	**Y**	n
126	Semarchy xDM	**Y**	**Y**	n	n	**Y**	**Y**	**Y**	n	**Y**	n	n	**Y**	**Y**	**Y**	**Y**	**Y**	n	n	n

Y - *"yes"*, n - *"no"*

Selection 2: Y - *"included"*, Y* - *"included as alternative"*, n - *"excluded"*

Continuation of Appendix IV

ID	Tool	Data Profiling	Custom DQ Rules	DQ Rule Definition in SQL	DQ Dimensions Used	DQ Rules Repository	Erroneous Records Shown	DQ Report Creation	DQ Dashboard	Data Match Detection	Anomaly Detection	DQ Rule Detection	Data Cleansing	Data Enrichment	Master Data Management	Data Lineage	Data Catalogue	Data Semantics discovery	Data Integration	Selection 2
130	SQL Power DQguru	n	**Y**	n	n	n	n	n	n	**Y**	n	n	**Y**	**Y**	n	n	n	n	n	n
131	Stratio Augmented Data Fabric Platform	n	**Y**	n	n	**Y**	n	**Y**	n	**Y**	n	n	**Y**	**Y**	n	**Y**	**Y**	**Y**	**Y**	n
133	Syniti Knowledge Platform	**Y**	**Y**	n	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**
136	RingLead Platform	n	n	n	n	n	n	n	n	n	n	n	**Y**	**Y**	n	n	n	n	n	n
137	ZoomInfo OperationsOS	n	n	n	n	n	n	n	n	**Y**	n	n	**Y**	**Y**	**Y**	n	n	n	n	n
138	Tale Of Data	**Y**	n	n	n	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	n	**Y**	**Y**	n	n	n	**Y**	n	n
139	Talend Data Fabric	**Y**	n	n	**Y**	n	n	n	**Y**	**Y**	**Y**	n	**Y**	**Y**	n	**Y**	**Y**	**Y**	**Y**	n
140	Talend Data Preparation	**Y**	**Y**	n	**Y**	n	n	n	n	n	n	n	**Y**	**Y**	n	n	n	n	n	n
141	Talend Data Stewardship	**Y**	**Y**	n	**Y**	**Y**	**Y**	**Y**	n	n	n	n	**Y**	n	n	n	n	n	n	n
144	TIBCO Clarity	**Y**	n	n	**Y**	n	**Y**	n	n	**Y**	n	n	**Y**	**Y**	n	n	n	n	n	n
146	iCEDQ	n	**Y**	n	n	**Y**	**Y**	**Y**	n	n	n	n	n	n	n	n	n	n	n	n
148	Uniserv	n	n	n	n	n	n	n	n	**Y**	n	n	**Y**	n	n	n	n	n	n	n

Y - *"yes"*, n - *"no"*

Selection 2: Y - *"included"*, Y* - *"included as alternative"*, n - *"excluded"*

Continuation of Appendix IV

ID	Tool	Data Profiling	Custom DQ Rules	DQ Rule Definition in SQL	DQ Dimensions Used	DQ Rules Repository	Erroneous Records Shown	DQ Report Creation	DQ Dashboard	Data Match Detection	Anomaly Detection	DQ Rule Detection	Data Cleansing	Data Enrichment	Master Data Management	Data Lineage	Data Catalogue	Data Semantics discovery	Data Integration	Selection 2
149	DataFuse	n	n	n	n	n	n	n	n	n	n	n	**Y**	**Y**	n	n	n	n	n	n
150	DemandTools	**Y**	n	n	n	n	n	n	n	n	n	n	**Y**	n	n	n	n	n	n	n
151	Clean & Match Enterprise	**Y**	n	n	n	n	n	n	n	**Y**	n	n	**Y**	n	n	n	n	n	n	n

Y - *"yes"*, n - *"no"*

Selection 2: Y - *"included"*, Y* - *"included as alternative"*, n - *"excluded"*

V. DQ Tools: Environment Features

ID	Tool	Tool Environment	Data Processing Environment	API Used	Flat file (.txt, .csv, .tsv)	API Used	Spreadsheet	JSON	Relational Database	Non-Relational Database	DW	Data Lake	Also in Cloud	Selection 3
6	MobyDQ	Cloud		n	n	n	n	**Y**	n	**Y**	n	**Y**	n	n
8	AbInitio Enterprise Data Platform	Cloud/Hybrid	Both	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	n	**Y**
12	Anomalo	Cloud	Both	n	n	n	n	**Y**	n	**Y**	**Y**	**Y**	n	Y*
16	Ataccama ONE	Cloud	Both	**Y**	**Y**	**Y**	**Y**	**Y**	n	**Y**	**Y**	**Y**	n	**Y**
21	Smart Data Platform	Cloud		n	n	n	n	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	n
25	Collibra Platform	Hybrid	Private Cloud	**Y**	**Y**	n	**Y**	**Y**	n	**Y**	**Y**	**Y**	**Y**	**Y**
39	DQLABS Platform	Cloud	Private Cloud	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**
41	DvSum	Cloud	Both	**Y**	**Y**	**Y**	n	**Y**	**Y**	**Y**	**Y**	**Y**	n	**Y**
44	Experian Aperture Data Studio	Cloud		n	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	n
52	Global IDs Data Quality Suites	Cloud/On-prem	Both	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**
68	Informatica Cloud Data Quality	Cloud	Private Cloud	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**
70	Informatica Data Engineering Quality	Cloud/On-prem	Both	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**
72	Informatica Master Data Management	Cloud	Private Cloud	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**

Y - *"yes"*, n - *"no"*

Selection 3: Y - *"included"*, Y* - *"alternative included"*, n - *"excluded"*

Continuation of Appendix V

ID	Tool	Tool Environment	Data Processing Environment	API Used	Flat file (.txt, .csv, .tsv)	API Used	Spreadsheet	JSON	Relational Database	Non-Relational Database	DW	Data Lake	Also in Cloud	Selection 3
85	LiTech Data Quality Management	On-prem	On-prem	**Y**	**Y**	n	n	**Y**	n	**Y**	**Y**	**Y**	**Y**	Y*
96	Monte Carlo Data Observability Platform	Cloud	Vendor Cloud	n	n	n	n	**Y**	n	**Y**	**Y**	**Y**	n	n
109	Precisely Data360	Cloud	Vendor Cloud	**Y**	**Y**	**Y**	**Y**	**Y**	n	**Y**	**Y**	**Y**	**Y**	n
110	Precisely Spectrum Quality			**Y**	**Y**	**Y**	**Y**	**Y**	n	**Y**	n	**Y**	n	n
119	SAP Information Steward	Cloud	Vendor Cloud	**Y**	**Y**	**Y**	n	**Y**	n	**Y**	n	**Y**	**Y**	n
133	Syniti Knowledge Platform	Cloud	Both	**Y**	n	n	n	**Y**	n	**Y**	n	**Y**	**Y**	**Y**

Y - *"yes"*, n - *"no"*

Selection 3: Y - *"included"*, Y* - *"alternative included"*, n - *"excluded"*

VI. DQ Tools: Descriptions of DQ Rule Detectors

ID	Tool	Description
8	AbInitio Enterprise Data Platform	AbInitio includes a wide range of capabilities as it is listed on their website[5]. It claims to have almost all functionalities presented in the review report in Table 2, except DQ rules expression in SQL and DQ dashboard. It also supports all data sources reviewed, including DWs, and provides data processing wherever the data is, e.g. on-premises, virtual private cloud. AbInitio has the main feature of the goal of this thesis, automated DQ rule detection. It generates and applies DQ rules based on metadata that have also been automatically detected. Specifically, there are used data fields' types and expected relationships in the data to generate the DQ rules. If the rules are based on data field types, i.e., text, integer, date, etc., then it may recognise the data format to some extent, but may not detect rules for detailed patterns. The relationship, like the "start date" and "end date" of something, may give input for rules of the consistency dimension. It is not clear if there is detected external consistency, the relationship between the attributes of different data objects. There is also no information if data lineage as one type of metadata is used for generating reconciliation rules for checking the DQ of transformation and loading processes. AbInitio has plenty of well-described capabilities and is able to provide the customer with quite individual products. However, the descriptions on websites include much marketing-flavoured text, the platform is not possible to try out, and there is also no publicly available documentation.

[5] https://www.abinitio.com/en/

Continuation of Appendix VI		
ID	Tool	Description
16	Ataccama ONE Platform	ONE Platform is the second most popular DQ tool, being brought out by eight (8) sources (see Appendix IIa). It is a combination of DQ, master data, data catalogue, reference data, etc. solutions, and designed for different roles of data specialists. It has all the features reviewed, except DQ rules representation in SQL. It supports almost every data source reviewed, except non-relational databases. It is claimed to be designed for big data management[6]. Ataccama ONE platform's DQ component is clearly a rule-based DQ solution, having automated DQ rules detection and an option for self-defined rules. Rules are complemented with dimension and the result can be seen on the business term, data element, and report level. In the demo video[7] it is shown how all the automated and self-defined rules can be configured and published. Automated DQ rules are created from a library of built-in rules, data domains and business terms detected using a self-learning machine learning method as said on Ataccama's webpage[8]. It is supplemented by anomaly detection which is basically a self-learning outlier detection. Based on the anomalies DQ rules can also be added manually. Unfortunately, the Ataccama ONE platform is not trialable. Instead, the data profiling tool Ataccama DQ Analyzer is only provided. It does not detect DQ rules or anomalies. Also, automated DQ rule generation of Ataccama ONE is not introduced in the demo video and there is no available documentation, but the author finds the user interface simple and logical. Visualising the data lineage looks also clear and understandable which is helpful for responsible roles for reporting the DQ of critical data elements in DWs.

[6] https://www.ataccama.com/platform
[7] https://www.youtube.com/watch?v=XG6n2CMGJ-4&t=10s
[8] https://www.ataccama.com/platform/data-quality

Continuation of Appendix VI		
ID	Tool	Description
25	Collibra Data Intelligence Cloud	Collibra is a full-service data platform, including services, like DQ, data catalogue, data governance, data lineage, etc. It has an available demo, a platform where can be signed in. There can be taken a guided tour which introduces connecting to some data sources, detecting anomalies with automatically generated rules, building custom rules, and monitoring the DQ. This DQ product is mapped to all DQ features in scope. It profiles data, detects anomalies, generates automatic DQ rules, and allows to define own DQ rules. It is possible to define DQ rule expressions in SQL. Collibra has metadata functionalities, works hybrid and can process data in the private cloud. It uses associative, unsupervised machine learning to auto-generate SQL-based, explainable and adaptive DQ rules. It creates snapshots and baselines to benchmark past data, constantly learns from new data and makes predictions for typos, formatting issues, outliers, relationships and more [Col]. In addition, Collibra has publicly available documentation[9] with many detailed guides and screenshots, and a short introductory video[10]. Overall, the author finds that the website, documentation and demo are well-described and visually easy to look at.
39	DQLABS Platform	DQLabs assembles DQ, metadata, and data governance functionalities, trying to bridge the gap between technical and business users as DQLabs has claimed on their website[11]. Their lead sentence is "Observe, Measure and Discover all in one platform". DQLabs uses machine learning for discovering and extracting semantics or business terms from a customer's data stack, identifying the data type and its sensitivity level (e.g., PII), and detecting anomalies in data. It is said that discovered metadata is used for the automated generation of DQ rules. There are supported all data sources which were determined in the review protocol: different files, relational and non-relational databases, DWs, and data lakes, and they provide also API. DQLabs does not provide immediately a demo or a free trial. It does not have any introductory or even marketing purposes video, and there is no publicly available documentation. The author requested a demo and was contacted by phone, but there was not given any additional information. It was called only for selling purposes, requesting more information from the author.

[9] https://productresources.collibra.com/docs/collibra/latest/Content/Home.htm
[10] https://www.youtube.com/watch?v=gsUM8lX8DHA
[11] https://www.dqlabs.ai/platform/

Continuation of Appendix VI		
ID	Tool	Description
41	DvSum	DvSum aggregates DQ, data catalogue and data governance functionalities for several data sources, including data lake or data warehouse. It additionally provides a chatbot functionality where users can chat with data asking questions about the data. An example question provided on their webpage[12], "How many of our therapy patients use Android vs iPhone?" The Chatbot is able to return a bar chart with respective counts of Android users, iPhone users and other users of therapy patients. On a high level, DvSum claims to have automated DQ checks, which is in the interest of this thesis, DQ monitoring and integration to pipelines, a self-service root-cause analysis of data issues, and the impact analysis of data model changes. DvSum suggests automatically AI-driven DQ checks that are combinations of statistical anomaly detection and rule-based algorithms. These checks are able to validate data types, empty values, volume, and shifts in data distribution[13]. DvSum has no publicly available documentation and the DQ solution is not immediately trialable. The data catalogue function can be tried out with a demo, but for the DQ solution, there has to be scheduled a call.
52	Global IDs Data Quality Suites	Global IDs DQ Suites is a part of Global IDs DEEP Platform[14] which comprises a set of core functions: automated discovery and profiling, data classification, data lineage, DQ, etc. Additionally, Global IDs can automatically generate the DQ controls for critical data elements. These controls function like rules that continuously monitor the data elements they are linked to. Its architecture is designed for integration from the ground up with all platform functionality accessible via APIs and claims to be able to automate data management for enterprises of any size or data ecosystem. They provide deployment on-premises or in the cloud and have a big library of connectors to several structured databases, applications, file types, NoSQL, Big Data, and the cloud. There is no immediately available demo or a free trial, and publicly available documentation. Nevertheless, they have a very interesting podcast[15] about the importance of data lineage. There is told that data lineage is important not only for regulative purposes but also for checking DQ and tracing issues. It is essential to know the lineage when making data reconciliation to be sure that data flows correctly through the pipelines. Global IDs do not reveal how their DQ controls generation work.

[12]https://dvsum.ai/

[13]https://dvsum.ai/solutions/data-quality/

[14]`https://www.globalids.com/platform-features/`

[15]`https://share.transistor.fm/s/16467431`

Continuation of Appendix VI		
ID	Tool	Description
		Informatica cloud-based products use the CLAIRE engine that delivers metadata-driven artificial intelligence to Informatica's cloud services, enabling intelligent recommendations of DQ rules that are based on how similar data has been managed prior [Infa] [Infb] [Infc] which may mean that CLAIRE learns from the previous "experience" of the specific user or rather actions of the community. All cloud-based Informatica services have connectors to a lot of data sources brought out in their website[16]. There are connectors for files, special applications, DWs, data lakes, databases, and even to social media platforms, like Facebook and LinkedIn. Informatica does not share its documentation in public for all the services, but these can be found on the website[17]. Informatica provides an interactive demo for the Master Data Management product, but the author finds it uninformative and inconvenient.
68	Informatica Cloud DQ	Informatica Cloud DQ is an AI-based DQ solution for small companies to big enterprises [Infa]. It is mapped to almost all DQ functionalities in the report, except defining DQ rules in SQL and data enrichment.
70	Informatica Data Engineering Quality	Informatica DEQ is a DQ solution for large enterprises, and for all of its business applications on-premises, in the cloud, or big data, including Hadoop, NoSQL, and other environments. It has features for all stakeholders, like line-of-business managers, business analysts, data stewards, and IT personnel [Infb].
72	Informatica Master Data Management	Informatica MDM is the SaaS solution with all-in-one capabilities for master data management[18], e.g. metadata management, DQ management, data governance, data modelling, data integration, etc.

[16] https://www.informatica.com/products/cloud-integration/connectivity/connectors.html
[17] https://docs.informatica.com/
[18] https://www.informatica.com/products/master-data-management.html

Continuation of Appendix VI		
ID	Tool	Description
133	Syniti Knowledge Platform	SKP is an AI-powered cloud data platform uniting DQ, data catalogue and data governance functionalities. Its key capabilities include automated metadata scanning and profiling, automatic data catalogue creation, rule recommendation engine, AI-driven data matching and deduplication solution, validation, cleansing and enriching capabilities, DQ reports, and executive-level business outcome dashboards [Syna]. Syniti does not publish how exactly their rule engine works. It is said only that it is using AI for different data management functionalities, but it does not mention if it uses automatically scanned metadata and data catalogue or something else, like built-in rules. Syniti Knowledge Platform connects to the following data sources: relative databases, DWs and applications listed in their website [19]. Syniti Knowledge Platform works in the cloud but processes the data on the customer side where Syniti Connector is set up for communication. Knowledge Platform sends out commands and gets back metadata and metrics [Synb].

[19] https://www.syniti.com/solutions/data-replication/data-replication-supported-databases/

VII. Licence

Non-exclusive licence to reproduce thesis and make thesis public

I, **Heidi Carolina Martinsaari**,
(author's name)

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

   reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,

   Toward an Automated DQ Rule Detection in DWs,
   (title of thesis)

   supervised by Anastasija Nikiforova.
   (supervisor's name)

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, 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. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

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

Heidi Carolina Martinsaari
09/05/2023