Journal of Big Data

RESEARCH Open Access

Big data quality framework: a holistic approach to continuous quality management

Ikbal Taleb[1], Mohamed Adel Serhani[2*], Chafik Bouhaddioui[3] and Rachida Dssouli[4]

*Correspondence: serhanim@uaeu.ac.ae
[2] College of Information Technology, UAE University, P.O. Box 15551, Al Ain, United Arab Emirates
Full list of author information is available at the end of the article

Abstract

Big Data is an essential research area for governments, institutions, and private agencies to support their analytics decisions. Big Data refers to all about data, how it is collected, processed, and analyzed to generate value-added data-driven insights and decisions. Degradation in Data Quality may result in unpredictable consequences. In this case, confidence and worthiness in the data and its source are lost. In the Big Data context, data characteristics, such as volume, multi-heterogeneous data sources, and fast data generation, increase the risk of quality degradation and require efficient mechanisms to check data worthiness. However, ensuring Big Data Quality (BDQ) is a very costly and time-consuming process, since excessive computing resources are required. Maintaining Quality through the Big Data lifecycle requires quality profiling and verification before its processing decision. A BDQ Management Framework for enhancing the pre-processing activities while strengthening data control is proposed. The proposed framework uses a new concept called Big Data Quality Profile. This concept captures quality outline, requirements, attributes, dimensions, scores, and rules. Using Big Data profiling and sampling components of the framework, a faster and efficient data quality estimation is initiated before and after an intermediate pre-processing phase. The exploratory profiling component of the framework plays an initial role in quality profiling; it uses a set of predefined quality metrics to evaluate important data quality dimensions. It generates quality rules by applying various pre-processing activities and their related functions. These rules mainly aim at the Data Quality Profile and result in quality scores for the selected quality attributes. The framework implementation and dataflow management across various quality management processes have been discussed, further some ongoing work on framework evaluation and deployment to support quality evaluation decisions conclude the paper.

Keywords: Big data quality, Data quality profile, Quality assessment, Quality metrics and scores, Pre-processing

Introduction

Big Data is universal [1], it consists of large volumes of data, with unconventional types. These types may be structured, unstructured, or in a continuous motion. Either it is used by the industry and governments or by research institutions, a new way to handle Big Data from a technology perspective to research approaches in its management is highly required to support data-driven decisions. The expectation from

Big Data analytics varies from trends finding to pattern discovery in different application domains such as healthcare, businesses, and scientific exploration. The aim is to extract significant insights and decisions. Extracting this precious information from large datasets is not an easy task. A devoted planning and appropriate selection of tools and techniques are available to optimize the exploration of Big Data.

Owning a huge amount of data does not often lead to valuable insights and decisions since Big Data does not necessarily mean Big insights. In fact, it can complicate the processes involved in fulfilling such expectations. Also, a lot of resources may be required, in addition to adapting the existing analytics algorithms to cope with Big Data requirements. Generally, data is not ready to be processed as it is. It should go through many stages, including cleansing and pre-processing, before undergoing any refining, evaluation, and preparation treatment for the next stages along its lifecycle.

Data Quality (DQ) is a very important aspect of Big Data for assessing the aforementioned pre-processing data transformations. This is because Big Data is mostly obtained from the web, social networks, and the IoT, where they may be found in a structured or unstructured form with no schema and eventually with no quality properties. Exploring data profiling, and more specifically, DQ profiling is essential before data preparation and pre-processing for both structured and unstructured data. Also, a DQ assessment should be conducted for all data-related content, including attributes and features. Then, an analysis of the assessment results can provide the necessary elements to enhance, control, monitor, and enforce the DQ along the Big Data lifecycle; for example, maintaining high Data Quality (conforming to its requirements) in the processing phase.

Data Quality has been an active and attractive research area for several years [2, 3]. In the context of Big Data, quality assessment processes are hard to implement, since they are time- and cost-consuming, especially for the pre-processing activities. These issues have got intensified since the available quality assessment techniques were developed initially for well-structured data and are not fully appropriate for Big Data. Consequently, new Data Quality processes must be carefully developed to assess the data origin, domain, format, and type. An appropriate DQ management scheme is critical when dealing with Big Data. Furthermore, Big Data architectures do not incorporate quality assessment practices throughout the Big Data lifecycle apart from pre-processing. Some new initiatives are still limited to specific applications [4–6]. However, the evaluation and estimation of Big Data Quality should be handled in all phases of the Big Data lifecycle from data inception to its analytics, thus support data-driven decisions.

The work presented in this paper is related to Big Data Quality management through the Big Data lifecycle. The objective of such a management perspective is to provide users or data scientists with a framework capable of managing DQ from its inception to its analytics and visualization, therefore support decisions. The definition of acceptable Big Data quality depends largely on the type of applications and Big Data requirements. The need for a quality Big Data evaluation before engaging in any Big Data related project is imminent. This is because the high costs involved in processing useless data at an early stage of its lifecycle can be prevented. More challenges to the data quality evaluation process may occur when dealing with unstructured, schema-less data collected from multiples sources. Moreover, a Big Data Quality Management Framework can provide quality management mechanisms to handle and ensure data quality throughout the Big Data lifecycle by:

- Improving the processes of the Big Data lifecycle to be quality-driven, in a way that it integrates quality assessment (built-in) at every stage of the Big Data architecture.
- Providing quality assessment and enhancement mechanisms to support cross-process data quality enforcement.
- Introducing the concept of Big Data Quality Profile (DQP) to manage and trace the whole data pre-processing procedures from data source selection to final pre-processed data and beyond (processing and analytics).
- Supporting profiling of data quality and quality rules discovery based on quantitative quality assessments.
- Supporting deep quality assessment using qualitative quality evaluations on data samples obtained using data reduction techniques.
- Supporting data-driven decision making based on the latest data assessments and analytics results.

The remainder of this paper is organized as follows. In Sect. "Overview and background", we provide ample detail and background on Big Data and data quality, besides, the introduction of the problem statement, and the research objectives. The research literature related to Big Data quality assessment approaches is presented in Sect. "Related research studies". The components of the proposed framework and an explanation of their main functionalities are described in Sect. "Big data quality management framework". Finally, implementation discussion and dataflow management are detailed in Sect. "Implementations: Dataflow and quality processes development", whereas Sect. "Conclusion" concludes the paper and points to our ongoing research developments.

Overview and background

Big data

An exponential increase in global inter-network activities and data storage has triggered the Big Data Era. Moreover, application domains, including Facebook, Amazon, Twitter, YouTube, Internet of Things Sensors, and mobile smartphones, are the main players and data generators. The amount of data generated daily is around 2.5 quintillion bytes (2.5 Exabyte, 1 EB = 1018 Bytes).

According to IBM, Big Data is a high-volume, high-velocity, and high-variety information asset that demands cost-effective, innovative forms of information processing for enhanced insights and decision-making. It is used to describe a massive volume of both structured and unstructured data; therefore, Big Data processing using traditional database and software tools is a difficult task. Big Data also refers to the technologies and storage facilities required by an organization to handle and manage large amounts of data.

Originally, in [7], the McKinsey Global Institute identifies three Big Data characteristics commonly known as ”3Vs” for Volume, Variety, and Velocity [1, 7–11]. These characteristics have been extended to more dimensions, moving to 10 Vs (Volume, Velocity, Variety, Veracity, Value, Vitality, Viscosity, Visualization, Vulnerability) [12–14].

In [10, 15, 16], the authors define important Big Data systems architectures. The data in Big Data comes from (1) heterogeneous data sources (e-Gov: Census data, Social networking: Facebook, and Web: Google page rank data), (2) data in different formats (video, text), and (3) data of various forms (unstructured: raw text data with no schema, and semi-structured: metadata, graph structure as text). Moreover, data travels through different stages, composing the Big Data lifecycle. Many aspects of Big Data architectures were compiled from the literature. Our enhanced design contributions are illustrated in Fig. 1 and described as follows:

- **Data generation:** this is the phase of data creation. Many data sources can generate this data such as electrophysiology signals, sensors used to gather climate information, surveillance devices, posts to social media sites, videos and still images, transaction records, stock market indices, GPS location, etc.
- **Data acquisition:** it consists of data collection, data transmission, and data preprocessing [1, 10]. Due to the exponential growth and availability of heterogeneous data production sources, an unprecedented amount of structured, semi-structured, and unstructured data is available. Therefore, the Big Data Pre-Processing consists of typical data pre-processing activities: *integration, enhancements and enrichment, transformation, reduction, discretization, and cleansing***.**
- **Data storage:** it consists of the data center infrastructure, where the data is stored and distributed among several clusters and data centers, spread geographically around the world. The software storage is supported by the Hadoop ecosystem to ensure a certain degree of fault tolerance storage reliability and efficiency through replication. The data storage stage is responsible for all input and output data that circulates within the lifecycle.
- **Data analysis:** *(Processing, Analytics, and Visualization);* it involves the application of data mining and machine learning algorithms to process the data and extract useful insights for better decision making. Data scientists are the most valuable users of this phase since they have the expertise to apply what is needed, on what must be analyzed.

Data quality, quality dimensions, and metrics

The majority of studies in the area of DQ originate from the database context [2, 3] and management research communities. According to [17], DQ is not an easy concept to

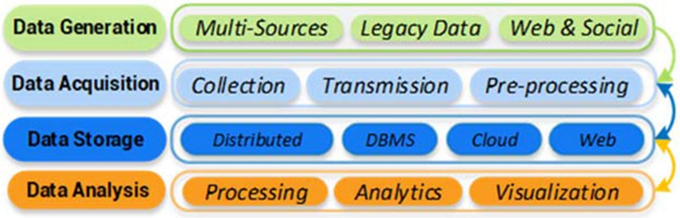


Fig. 1 Big data lifecycle value chain

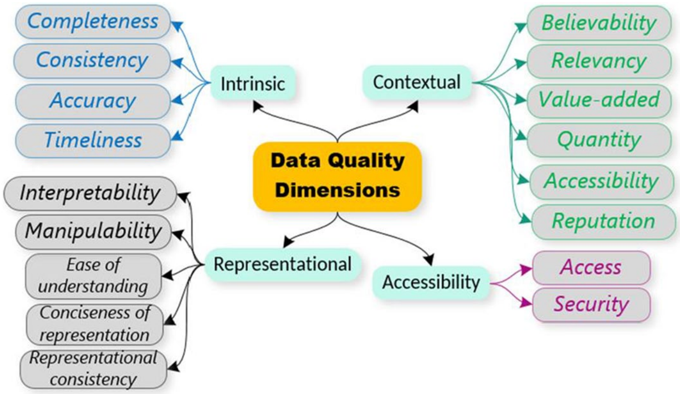


Fig. 2 Data quality dimensions

Table 1 Example of data quality dimensions (Intrinsic)

DQD's	Description
Completeness	Describes whether all relevant data is recorded. It measures missing values for an attribute
Consistency	Checks whether data agrees with its format and structure. It mostly refers to the respect of data constraints
Accuracy	Measures whether data was recorded correctly and reflects realistic values. It is also defined as the "closeness of the agreement between the result of a measurement and a true value of the measure". [29]
Timeliness	Computes whether data is up to date, referred to as data currency and volatility. [30]

define. Its definition is data domain awareness. There is a consensus that data quality always depends on the quality of the data source [18]. However, it highlights that enormous quality issues are hidden inside data and their values.

In the following, the definitions of data quality, data quality dimensions, and quality metrics and their measurements are given:

- **Data quality:** It has many meanings that are related to data context, domain, area, and the fields from which it is used [19, 20]. Academia interprets DQ differently than industry. In [21], data quality is reduced to "The capability of data to satisfy stated and implied needs when used under specified conditions". Also, DQ is defined as "fitness for use". Yet, [20] define data quality as the property corresponding to quality management, which is appropriate for use or meeting user needs.
- **Data quality dimensions:** DQD's are used to measure, quantify, and manage DQ [20, 22, 23]. Each quality dimension has a specific metric, which measures its performance. There are several DQDs, which can be organized into 4 categories according to [24, 25], intrinsic, contextual, accessibility, and representational [14, 15, 22, 24, 26, 27]. Two important categories (intrinsic and contextual) are illustrated in Fig. 2. Examples of intrinsic quality dimensions are illustrated in Table 1.
- **Metrics and measurements:** Once the data is generated, its quality should be measured. This means that a data-driven strategy is considered to act on the data. Hence, it is mandatory to measure and quantify the DQD. Structured or semi-

Table 2 – Example of data quality metrics

DQD's	Metric functions
Completeness	Comp(%) = **NNMV*100/N**: Number of non-missing values /N
Consistency	Cons(%) = **NVRC*100/N**: Number of values that respects constraints /N
Accuracy	Acc(%) = **NCV*100/N**: Number of correct values /N
Uniqueness	Uniq(%) = **NDV*100/N**: Number of distinct values /N
N	Total number of observations (Rows) in dataset or sample

Table 3 Example of intrinsic DQD's vs. big data characteristics

Data quality dimensions	Big Data V's			
	Volume	Velocity	Variety	Veracity
Accuracy	X		X	X
Completeness	X	X		X
Consistency		X	X	X
Currency		X		X

structured data is available as a set of attributes represented in columns or rows, and their values are respectively recorded. In [28], a quality metric, as a quantitative or categorical representation of one or more attributes, is defined. Any data quality metric should define whether the values of an attribute respect a targeted quality dimension. The author [29], quoted that data quality measurement metrics tend to evaluate binary results: correct or incorrect, or a value between 0 and 100 (with 100% representing the highest). This applies to some quality dimensions such as accuracy, completeness, consistency, and currency. Examples of DQD metrics are illustrated in Table 2.

DQD's must be relevant to data quality problems that have been identified. Thus, a metric tends to measure if attributes comply with defined DQD's. These measurements are performed for each attribute, given their type and data ranges of values collected from the data profiling process. The measurements produce DQD's scores for the designed metrics of all attributes [30]. Specific metrics need to be defined, to estimate specific quality dimensions of other data types such as images, videos, and audio [5].

Big data characteristics and data quality

The main Big Data characteristics, commonly named as V's, are initially, Volume, Velocity, Variety, and Veracity. Since the Big Data inception, 10 V's have been defined, and probably new Vs will be adopted [12]. For example, veracity tends to express and describe the trustworthiness of data, mostly known as data quality. The accuracy is often related to precision, reliability, and veracity [31]. Our tentative mapping among these characteristics, data, and data quality, is shown in Table 3. It is based on the intuitive studies accomplished by [5, 32, 33]. In these studies, the authors attempted to link the V's to the data quality dimensions. In another study, the authors [34] addressed the mapping of DQD Accuracy with the Big Data characteristic Volume and showed that the data size has an impact on DQ.

Big data lifecycle: where quality matters?

According to [21, 35], data quality issues may appear in each phase of the Big Data value chain. Addressing data quality may follow different strategies, as each phase has its features either improving the quality of existing data or/and refining, reassessing, redesigning the whole processes, which generate and collect data, aiming at improving their quality.

Big Data quality issues were addressed by many studies in the literature [36–38]. These studies generally elaborated on the issues and proposed generic frameworks with no comprehensive approaches and techniques to manage quality across the Big Data lifecycle. Among these, generic frameworks are presented in [5, 39, 40].

In Fig. 3, it is illustrated where data quality can and must be addressed in the Big Data value chain phases/stages from (1) to (7).

1. In the data generation phase, there is a need to define how and what data is generated.
2. In the data transmission phase, the data distribution scheme relies on the underlying networks. Unreliable networks may affect data transfer. Its quality is expressed by data loss and transmission errors.
3. Data collection refers to where, when, and how the data is collected and handled. Well-defined structured constraint verification on data must be established.
4. The pre-processing phase is one of the main focus points of the proposed work. It follows a data-driven strategy, which is largely focused on data. An evaluation process provides the necessary means to ensure the quality of data for the next phases. An evaluation of the DQ before (Pre) and after (Post) pre-processing on data samples is necessary to strengthen the DQP.
5. In the Big Data storage phase, some aspects of data quality, such as storage failure, are handled by replicating data on multiple storages. The latter is also valid for data transmission when a network fails to transmit data.
6. In the Data Processing and Analytics phases, the quality is influenced by both the applied process and data quality itself. Among the various data mining and machine learning algorithms and techniques suitable for Big Data, those that converge rapidly and consume fewer cloud resources will be highly adopted. The relation between DQ

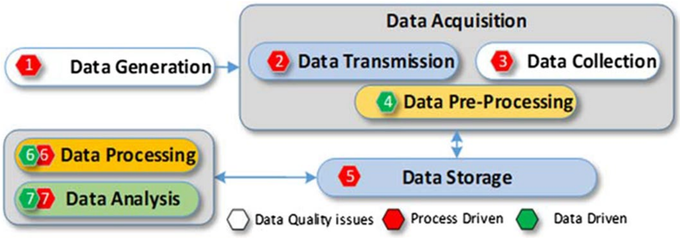


Fig. 3 Where quality matters in big data lifecycle?

and the processing methods is substantial. A certain DQ requirement on these methods or algorithms might be imposed to ensure efficient performance.

7. Finally, for an ongoing iterative value chain, the visualization phase seems to be only a representation of the data in a fashionable way such as a dashboard. This helps the decision-makers to have a clear picture of the data and its valuable insights. Finally, in this work, Big Data is transformed into useful Small Data, which is easy to visualize and interpret.

Data quality issues

Data quality issues generally appear when the quality requirements are not met on the data values [41]. These issues are due to several factors or processes having occurred at different levels:

(a) Data source level: unreliability, trust, data copying, inconsistency, multi-sources, and data domain.
(b) Generation level: human data entry, sensors' readings, social media, unstructured data, and missing values.
(c) Process level (acquisition: collection, transmission).

In [21, 35, 42], many causes of poor data quality were enumerated, and a list of elements, which affect the quality and DQD's was produced. This list is illustrated in Table 4.

Related research studies

Research directions on Big Data differ between industry and academia. Industry scientists mainly focus on the technical implementations, infrastructures, and solutions for Big Data management, whereas researchers from academia tackle theoretical issues of Big Data. Academia's efforts mainly include the development of new algorithms for data analytics, data replication, data distribution, and optimization of data handling. In this section, the literature review is classified into 3 categories, which are described in the following sub-sections.

Data quality assessment approaches

Existing studies on data quality have been approached from different perspectives. In the majority of the papers, the authors agree that data quality is related to the phases or processes of its lifecycle [8]. Specifically, data quality is highly related to the data generation phases and/or with its origin. The methodologies adopted to assess data quality are based on traditional data strategies and should be adapted to Big Data. Moreover, the application domain and type of information (Content-based, Context-based, or Rating-based) affects the way the quality evaluation metrics are designed and applied. In content-based quality metrics, the information itself is used as a quality indicator, whereas in context-based metrics meta-data is used as quality indicators.

There are two main strategies to improve data quality according to [20, 23]: data-driven and process-driven. The first strategy handles the data quality in the pre-processing

Table 4 Example of data quality issues

		Data quality issues vs data quality dimensions (DQD)	**Accuracy**	**Completeness**	**Consistency**
Single data source	Cell instance level	Missing data	●	●	
		Incorrect data and references, Data entry errors and Misspelling	●		
		Irrelevant data			●
		Outdated data	●		
		Misfielded and contradictory values	●	●	●
	Dataset schema level	Domain and Uniqueness constrains, Functional dependency violation	●		
		Wrong data type, poor schema design			●
		Referential integrity violation, lack of integrity constraints	●	●	●
Multiple data source	Cell instance level	Different units, representations, Structural conflicts			●
		Different aggregation levels, Inconsistent aggregation	●		●
		Temporal mismatch, inconsistent timing	●		
	Dataset schema level	Heterogeneous data models and schema design	●	●	●
		Different encoding formats			●

phase by applying some pre-processing activities (PPA) such as cleansing, filtering, and normalization. These PPAs are important and occur before the data processing stage, preferably as early as possible. However, the process-driven quality strategy is applied to each stage of the Big Data value chain.

Data quality assessment was discussed early in the literature [10]. It is divided into two main categories: subjective and objective. Moreover, an approach that combines these two categories to provide organizations with usable data quality metrics to evaluate their data was proposed. However, the proposed approach was not developed to deal with Big Data.

In summary, Big Data quality should be addressed early in the pre-processing stage during the data lifecycle. The aforementioned Big Data quality challenges have not been investigated in the literature from all perspectives. There are still many open issues, which must be addressed especially at the pre-processing stage.

Rule-based quality methodologies

Since the data quality concept is context-driven, it may differ from an application domain to another. The definition of quality rules involves establishing a set of constraints on data generation, entry, and creation. Poor data can always exist, and rules are created or discovered to correct or eliminate this data. Rules themselves are only one part of the data quality assessment approach. The necessity to establish a consistent process for creating, discovering, and applying the quality rules should consider the following:

- Characterize the quality of data being good or bad from its profile and quality requirements.
- Select the data quality dimensions that apply to the data quality assessment context.
- Generate quality rules based on data quality requirements, quantitative, and qualitative assessments.
- Check, filter, optimize, validate, run, and test rules on data samples for efficient rules' management.
- Generate a statistical quality profile with quality rules. These rules represent an overview of successful valid rules with the expected quality levels.

Hereafter, the data quality rules are discovered from data quality evaluation. These rules will be used in Big Data pre-processing activities to improve the quality of data. The discovery process reveals many challenges, which should consider different factors, including data attributes, data quality dimensions, data quality rules discovery, and their relationship with pre-processing activities.

In (Lee et al., 2003), the authors concluded that the data quality problems depend on data, time, and context. Quality rules are applied to the data to solve and/or avoid quality problems. Accordingly, quality rules must be continuously assessed, updated, and optimized.

Most studies on the discovery of data quality rules come from the database community. These studies are often based on conditional functional dependencies (CFDs) to detect inconsistencies in data. CFDs are used to formulate data quality rules, which are generally expressed manually and discovered automatically using several CFD approaches [3, 43].

Data quality assessment in Big Data has been addressed in several studies. In [32], a Data Quality-in-Use model was proposed to assess the quality of Big Data. Business rules for data quality are used to decide on which data these rules must meet the pre-defined constraints or requirements. In [44], a new quality assessment approach was introduced and involved both the data provider and the data consumer. The assessment was mainly based on data consistency rules provided as metadata.

The majority of research studies on data quality and discovery of data quality rules are based on CFD's and database. In Big Data quality, the size, variety, and veracity of data are key characteristics that must be considered. These characteristics should be processed to reduce the quality assessment time and resources since they are handled before the pre-processing phase. Regarding quality rules, it is fundamental to consider these rules to eliminate poor data and enforce quality on existing data, while following a data-driven quality context.

Big data pre-processing frameworks

The pre-processing of data before performing any analytics is primeval. However, several challenges have emerged at this crucial phase of the Big Data value chain [10]. Data quality is one of these challenges, which must be highly considered in the Big Data context.

As pointed out in [45], data quality problems arise when dealing with multiple data sources. This increases the requirements for data cleansing significantly. Additionally, the large size of datasets, which arrive at an uncontrolled speed, generates an overhead on the cleansing processes. In [46–48], NADEEF, an extensible data cleaning system, was proposed. The extension for Big Data cleaning based on NADEEF was presented in [49] for streaming data. The system deals with data quality from the data cleaning activity using data quality rules and functional dependencies rules [14].

Numerous other studies on Big Data management frameworks exist. In these studies, the authors surveyed and proposed Big Data management models dealing with storage, pre-processing, and processing [50–52]. An up-to-date review of the techniques and methods for each process involved in the management processes is also included.

The importance of quality evaluation in Big Data Management has not been, generally, addressed. In some studies, Big Data characteristics are the only recommendations for quality. However, no mechanisms have been proposed to map or handle quality issues that might be a consequence of these Big Data Vs. A Big Data Management Framework, which includes data quality management, must be developed to cope with end-to-end quality management across the Big Data lifecycle.

Finally, it is worth mentioning that research initiatives and solutions on Big Data quality are still in their preliminary phase; there is much to do on the development and standardization of Big Data quality. Big Data quality is a multidisciplinary, complex, and multi-variant domain, where new evaluation techniques, processing and analytics algorithms, storage and processing technologies, and platforms will play a key role in the development and maturity of this active research area. We anticipate that researchers from academia will contribute to the development of new Big Data quality approaches, algorithms, and optimization techniques, which will advance beyond the traditional approaches used in databases and data warehouses. Additionally, industries will lead development initiatives of new platforms, solutions, and technologies optimized to support end-to-end quality management within the Big Data lifecycle.

Big data quality management framework

The purpose of proposing a Big Data Quality Management Framework (BDQMF) is to address the quality at all stages of the Big Data lifecycle. This can be achieved by managing data quality before and after the pre-processing stage while providing feedback at each stage and loop back to the previous phase, whenever possible. We also believe that data quality must be handled at data inception. However, this is not considered in this work.

To overcome the limitations of the existing Big Data architectures for managing data quality, a Big Data Quality pre-processing approach is proposed: a Quality Framework [53]. In our framework, the quality evaluation process tends to extract the actual quality status of Big Data and proposes efficient actions to avoid, eliminate, or enhance poor data, thus improving its quality. The framework features the creation and management of a DQP and its repository. The proposed scheme deals with data quality evaluation before and after the pre-processing phase. These practices are essential to ensure a certain quality level for the next phases while maintaining the optimal cost of the evaluation.

In this work, a quantitative approach is used. This approach consists of an end-to-end data quality management system that deals with DQ through the execution of pre-pre-processing tasks to evaluate BDQ on data. It starts with data sampling, data and DQ profiling, and gathering user DQ requirements. It then proceeds to DQD evaluation and discovery of Quality rules from quality scores and requirements. Each data quality rule is represented by one-to-many Pre-Processing Functions (PPF's) under a specific Pre-Processing Activity (PPA). A PPA, such as cleansing, aims at increasing data quality. Pre-processing is applied to Big Data samples and re-evaluated once again to update and certify that the quality profile is complete. It is applied to the whole Big Dataset, not only to data samples. Before pre-processing, the DQP is tuned and revisited by quality experts for endorsement based on an equivalent data quality report. This report states the quality scores of the data, not the rules.

Framework description

The BDQM framework is illustrated in Fig. 4, where all the components cooperate, relying on the Data Quality Profile. It is initially created as a Data Profile and is progressively extended from the data collection phase to the analytics phase to capture important quality-related information. For example, it contains quality requirements, targeted data quality dimensions, quality scores, and quality rules.

Data lifecycle stages are part of the BDQMF. Generated feedbacks in all the stages are analyzed and used to correct, improve the data quality, and detect any DQ management related failures. The key components of the proposed BDQMF include:

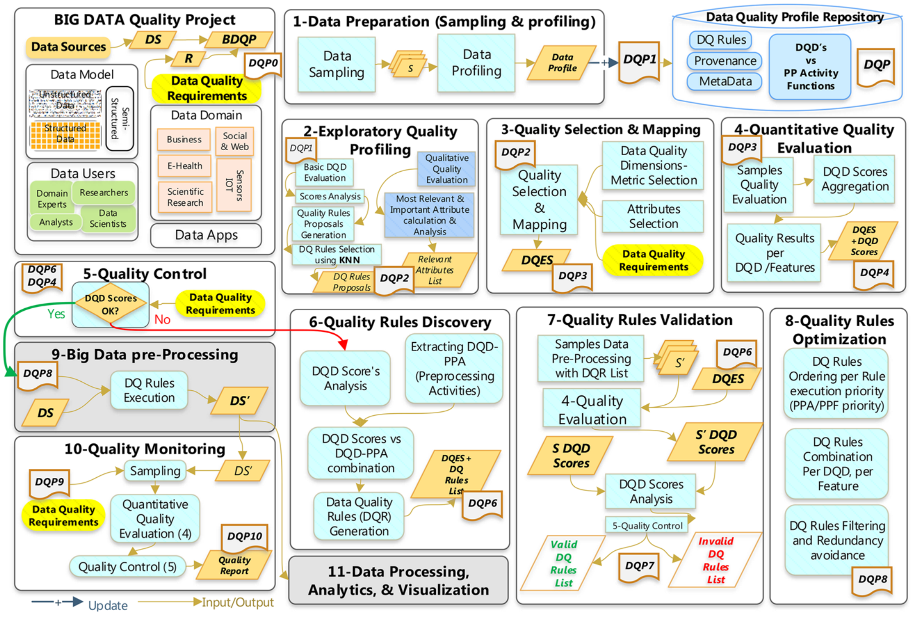

BIG DATA Quality Project
Data Sources
DS
BDQP
R
DQP0
Data Quality Requirements
Data Model
Unstructured Data
Structured Data
Semi-Structured
Data Domain
Business
Social & Web
E-Health
Scientific Research
Sensors IOT
Data Users
Domain Experts
Researchers
Analysts
Data Scientists
Data Apps
1-Data Preparation (Sampling & profiling)
Data Sampling
S
Data Profiling
Data Profile
DQP1
Data Quality Profile Repository
DQ Rules
Provenance
MetaData
DQD's vs PP Activity Functions
DQP
2-Exploratory Quality Profiling
DQP1
Basic DQD Evaluation
Qualitative Quality Evaluation
Scores Analysis
Most Relevant & Important Attribute calculation & Analysis
Quality Rules Proposals Generation
DQ Rules Selection using KNN
Relevant Attributes List
DQ Rules Proposals
DQP2
3-Quality Selection & Mapping
DQP2
Quality Selection & Mapping
Data Quality Dimensions-Metric Selection
Attributes Selection
DQES
DQP3
Data Quality Requirements
4-Quantitative Quality Evaluation
DQP3
Samples Quality Evaluation
DQD Scores Aggregation
Quality Results per DQD /Features
DQES +DQD Scores
DQP4
5-Quality Control
DQP6
DQP4
DQD Scores OK?
Data Quality Requirements
Yes
No
6-Quality Rules Discovery
DQD Score's Analysis
Extracting DQD-PPA (Preprocessing Activities)
DQD Scores vs DQD-PPA combination
DQES + DQ Rules List
Data Quality Rules (DQR) Generation
DQP6
7-Quality Rules Validation
Samples Data Pre-Processing with DQR List
S'
DQP6
DQES
4-Quality Evaluation
S DQD Scores
S' DQD Scores
DQD Scores Analysis
5-Quality Control
Valid DQ Rules List
DQP7
Invalid DQ Rules List
8-Quality Rules Optimization
DQ Rules Ordering per Rule execution priority (PPA/PPF priority)
DQ Rules Combination Per DQD, per Feature
DQ Rules Filtering and Redundancy avoidance
DQP8
9-Big Data pre-Processing
DQP8
DQ Rules Execution
DS
DS'
10-Quality Monitoring
Sampling
DS'
DQP9
Data Quality Requirements
Quantitative Quality Evaluation (4)
DQP10
Quality Control (5)
Quality Report
11-Data Processing, Analytics, & Visualization
Update
Input/Output

Fig. 4 Big data quality management framework

(1) Big Data Quality Project (Data Sources, Data Model, User/App Quality Requirements, Data domain),
(2) Data Quality Profile and its Repository,
(3) Data Preparation (Sampling and Profiling),
(4) Exploratory Quality Profiling,
(5) Quality Parameters and Mapping,
(6) Quantitative Quality Evaluation,
(7) Quality Control,
(8) Quality Rules Discovery,
(9) Quality Rules Validation,
(10) Quality Rules Optimization,
(11) Big Data Pre-Processing,
(12) Data Processing,
(13) Data Visualization, and
(14) Quality Monitoring.

A detailed description of each of these components is provided hereafter.

Framework key components

In the following sub-sections, each component is described. Its input(s) and output(s), its main functions, and its roles and interactions with the other framework's components, are also described. Consequently, at each Big Data stage, the Data Quality Profile is created, updated, and adapted until it achieves the quality requirements already set by the users or applications at the beginning of the Big Data Quality Project.

Big data quality project module

The Big Data Quality Project Module contains all the elements that define the data sources, and the quality requirements set by either the Big Data users or Big Data applications to represent the quality foundations of the Big Data project. As illustrated in Error! Reference source not found., any Big Data Quality Project should specify a set of quality requirements as targeted quality goals (Fig. 5).

It represents the first module of the framework. The Big Data quality project represents the starting point of the BDQMF, where specifications of the data model, data sources, and targeted quality goals for DQD and data attributes are defined. These requirements are represented as data quality scores/ratios, which express the acceptance level of the evaluated data quality dimensions. For example, 80% of data accuracy, 60% data completeness, and 85% data consistency are judged by quality experts as accepted levels (or tolerance ratios). These levels can be relaxed using a range of values, depending on the context, the application domain, and the targeted processing algorithm's requirements.

Let us denote by ***BDQP(DS, DS', Req)*** a Big Data Quality Project Request that initiates many automatic processes:

(1) A data sampling and profiling process.
(2) An exploratory quality profiling process, which is included in many quality assessment procedures.

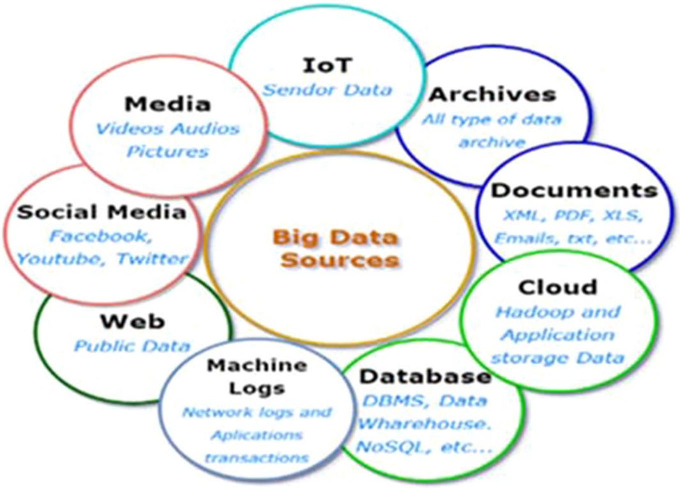


Fig. 5 Big data sources

(3) A pre-processing phase is eventually considered if the resulted quality scores are not met.

The ***BDQP*** contains the input dataset ***DS***, output dataset ***DS'***, and ***Req***. The Quality requirements are presented as a tuple of sets ***Req*** = (***D***, ***L***, ***A***), *where:*

- ***D*** represents a set of data quality dimensions DQD's (e.g., accuracy, consistency): $D = \{\boldsymbol{d}_0, \ldots, \boldsymbol{d}_{\boldsymbol{i}}, \ldots, \boldsymbol{d}_{\boldsymbol{m}}\}$,
- ***L*** is a set of DQD acceptance (tolerance) level ratios (%) set by the user or the application related to the quality project and associated with each DQD, respectively: $L = \{\boldsymbol{l}_0, \ldots, \boldsymbol{l}_{\boldsymbol{i}}, \ldots, \boldsymbol{l}_{\boldsymbol{m}}\}$,
- ***A*** is the set of targeted data attributes. If it is not specified, the DQD's are assessed for the dataset, which includes all possible attributes, since some dimensions need more detailed requirements to be assessed. Therefore, it depends on the DQD and the attribute type: $A = \{\boldsymbol{a}_0, \ldots, \boldsymbol{a}_{\boldsymbol{i}}, \ldots, \boldsymbol{a}_{\boldsymbol{m}}\}$

The Data quality requirements might be updated with some more aspects, whereas the profiling component provides well-detailed information about the data (**DQP Level 0**). This update is performed within the quality mapping component and interfaces with user experts to refine, reconfirm, and restructure their data quality parameters over the data attributes.

(a) **Data sources:** There are multiple Big Data sources. Most of them are generated from the new media (e.g., social media) based on the internet. Other data sources are based on the context of new technologies such as the cloud, sensors, and IoT. A list of Big Data sources is illustrated in Error! Reference source not found.

(b) **Data users, data applications, and quality requirements:** This module identifies and specifies the input sources of the quality requirements parameters for the data sources. These sources include user's quality requirements (e.g., Domain Experts, Researchers, Analysts, and Data scientists) or application quality requirements. (Applications may vary from simple data processing to machine learning applications or AI-based applications). For the users, a dashboard-like interface is used to capture user's data requirements and other quality information. This interface can be enriched with information from the data sources as attributes and their types, if available. This can efficiently guide users to the inputs and ensure the right data is used. This phase can be initiated after sample profiling or exploratory quality profiling. Otherwise, a general quality request is entered in the form of targeted Data Quality dimensions and their expected quality scores after the pre-processing phase. All the quality requirements parameters and settings are recorded in the Data Quality Profile (**DQP 0**). DQP Level 0 is created when the quality project is set.

The quality requirements are specifically set as quality score ratios, goals, or targets to be achieved by the BDQMF. They are expressed as targeted DQDs in the Big Data Quality Project.

Let us denote by ***Req***, a set of quality requirements presented as ***Req*** = $\{\boldsymbol{r}_0, \ldots, \boldsymbol{r}_{\boldsymbol{i}}, \ldots, \boldsymbol{r}_{\boldsymbol{m}}\}$ and constructed with a tuple (***D***, ***L, A***). The ***Req*** quality requirements list is identified by elements, where each of these elements is a quality requirement characterized by$\boldsymbol{r}_{\boldsymbol{i}} = (\boldsymbol{d}_{\boldsymbol{i}}, \boldsymbol{l}_{\boldsymbol{i}}, \boldsymbol{a}_{\boldsymbol{i}})$; $\boldsymbol{r}_{\boldsymbol{i}}$ represents a $\boldsymbol{d}_{\boldsymbol{i}}$ in the DQD with a minimum accepted ratio level $\boldsymbol{l}_{\boldsymbol{i}}$ for all or a sub-list of selected attributes $\boldsymbol{a}_{\boldsymbol{i}}$.

The initial DQP originating from this module is a DQP Level 0, containing the following tuple, as illustrated in Fig. 6: ***BDQP (DS, DS', Req) with Req*** = (***D***, ***L, A***)

(c) **Data models and data domains**

- **Data models:** If the Data is structured, then a schema is provided to add more detailed quality settings for all attributes. In other cases, if there are no such attributes or types, the data is considered as unstructured data, and its quality evaluation will consist of a set of general Quality Indicators (QI). In our Framework, these QI are provided especially for the cases, where a direct identification of DQD's is not available for an easy quality assessment.
- **Data domains:** Each data domain has a unique set of default quality requirements. Some are very sensitive to accuracy and completeness; others, prioritize data currency and higher timeliness. This module adds value to users or applications when it comes to quality requirements elicitation.

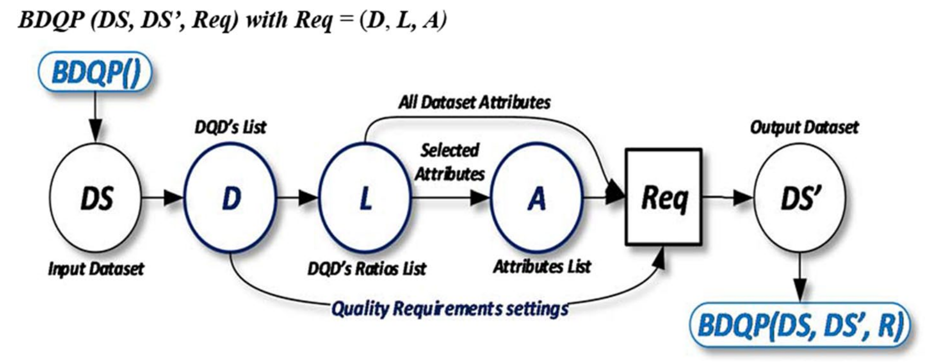


Fig. 6 BDQP and quality requirements settings

(d) **Data quality profile creation:** Once the Big Data Quality Project (BDQP) is initiated, the DQP level 0 (DQP0) is created and consists of the following elements, as illustrated in Fig. 7:

- Data sources information, which may include datasets, location, URL, origin, type, and size.
- Information about data that can be created or extracted from metadata if available, such as database schema, data attributes names and types, data profile, or basic data profile.
- Data domains such as business, health, commerce, or transportation.
- Data users, which may include the names and positions of each member of the project, security credentials, and data access levels.
- Data application platforms, software, programming languages, or applications that are used to process the data. These may include R, Python, Java, Julia, Orange, Rapid Miner, SPSS, Spark, and Hadoop.
- Data quality requirements: for each dataset, its expected quality ratios, and tolerance levels are accepted; otherwise, the data is discarded or repaired. It can also be set as a range of quality tolerance levels. For example, the DQD completeness is defined as equal to or higher than 67%, which means the acceptance ratio of missing values, is equal to or less than 33% (100% –67%).

Data quality profile (DQP) and repository (DQPREPO)

We describe hereafter the content of DQP and the DQP repository and the DQP levels captured through the lifecycle of framework processes.

a **Data quality profile**

1 The data quality profile is generated once a Big Data Quality Project is created. It contains, for example, information about the data sources, domain, attributes, or features. This information may be retrieved from metadata, data provenance,

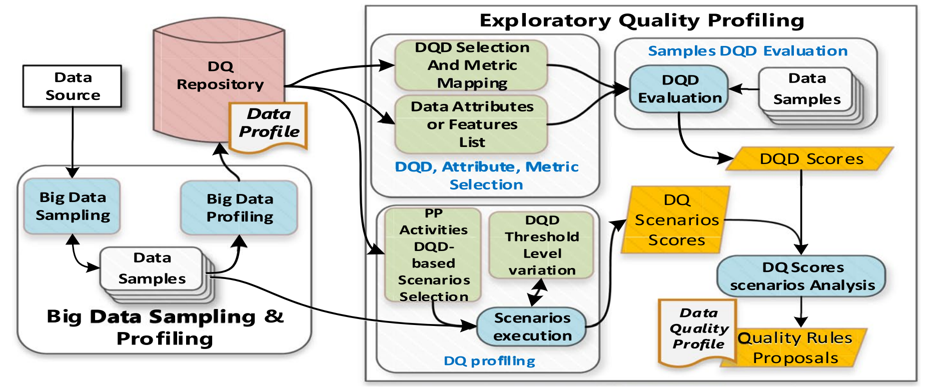

Exploratory Quality Profiling
Data Source
DQ Repository
Data Profile
Big Data Sampling
Big Data Profiling
Data Samples
Big Data Sampling & Profiling
DQD Selection And Metric Mapping
Data Attributes or Features List
DQD, Attribute, Metric Selection
Samples DQD Evaluation
DQD Evaluation
Data Samples
DQD Scores
PP Activities DQD-based Scenarios Selection
DQD Threshold Level variation
Scenarios execution
DQ profiling
DQ Scenarios Scores
DQ Scores scenarios Analysis
Data Quality Profile
Quality Rules Proposals

Fig. 7 Exploratory quality profiling modules

schema, or from the dataset itself. If not available, data preparation (sampling and profiling) is needed to collect and extract important information, which will support the upcoming processes, as the Data Profile (DP) is created.

2 An Exploratory Quality Profiling will generate a quality rules proposal list. The DP is updated with these rules and converted into a DQP. This will help the user to obtain an overview of some DQDs and make better attributes selection based on this first quality approximation with a ready-to-use list of rules for pre-processing.

3 The User/App quality requirements (Quality tolerance levels, DQDs, and targeted attributes) are set and added to the DQP. Updated and tuned-up previously proposed quality rules are more likely, or a complete redefinition of the quality requirement parameters is performed.

4 The mapping and selection phase will update the DQP with a DQES, which contains the set of attributes to be evaluated for a set of DQDs, using a set of metrics from the DQP repository.

5 The Quantitative Quality Evaluation component assesses the DQ and updates the DQES with DQD Scores.

6 The DQES scores pass through quality control if validated. The DQP is executed in the pre-processing stage and confirmed in the repository.

7 If the scores (based on the quality requirements) are not valid, a quality rules discovery, validation, and optimization will be added/updated to the DQP configuration to obtain a valid DQD score that satisfies the quality requirements.

8 A continuous quality monitoring is performed for an eventual DQ failure that triggers a DQP update.

(b) The DQP Repository: The ***DQPREPO*** contains detailed data quality profiles per data source and dataset. In the following, an information list managed by the repository is presented:

- Data Quality User/App requirements.
- Data Profiles, Metadata, and Data Provenance.
- Data Quality Profiles (e.g. Data Quality Evaluation Schemes, and Data Quality Rules).
- Data Quality Dimensions and related Metrics (metrics formulas and aggregate functions).
- Data Domains (DQD's, BD Characteristics).
- DQD's vs BD Characteristics.
- Pre-processing Activities (e.g. Cleansing, and Normalizing) and functions (to replace missing values).
- DQD's vs DQ Issues vs PPF: Pre-processing Functions.
- DQD's priority processing in Quality Rules.

At every stage, module, task, or process, the DQP repository is incrementally updated with quality-related information. This includes, for example, quality requirements, DQES, DQD scores, data quality rules, Pre-Processing activities, activity functions, DQD metrics, and Data Profiles. Moreover, the DQP's are organized per Data Domain and datatype to allow reuse. Adaptation is performed in the case of additional Big Datasets.

In Table 5, an example of DQP Repository managed information along with its pre-processing activities (PPA) and their related functions (PPAF), is presented.

(c) **DQP lifecycle (Levels)**: The DQP goes through the complete process flow of the proposed BDQMF. It starts with the specification of the Big Data Quality Project and ends with quality monitoring as an ongoing process that closes the quality enforcement loop and triggers other processes, which handle DQP adaptation, upgrade, or reuse. In Table 6, the various DQP levels and their interaction within the BDQM Framework components are described. Each component involves process operations applied to the DQP.

Data preparation: sampling and profiling

Data preparation generates representative Big Data samples that serve as an entry for profiling, quality evaluation, and quality rules validation.

Table 5 DQD's and their related pre-processing activities and functions

PPAF#	DQD	Metric	Data Type	Methods	Results (%)	PPA	PPAF	PPAF Related Actions or Proposals
11	Accuracy/ validity	Outliers detection	Num	Rule-based	Outliers Count/ Total Rows, List of Obs. with Outliers (Anomaly, Novelty)	Data cleansing	Retention	Use robust classification methods
12				Linear regression model			Winsorizing (Dealing with Outliers)	Replace outliers with closest values
13				High dimensional outlier detection methods			Exclusion, Truncation	Remove related rows
21	Completeness	Available data observation	All	Count the number of not (NA, Null, or any other values that express the Not Availability)	Not NA Count /Total observations (Rows)	Data enrichment	Data correction	Replace with mean
22								Replace with mode
23								Replace with median
24							Data removal	Remove Rows
25								Remove columns
26								Remove rows and cols

Table 6 Data quality profile levels

#	DQP Operation	Description	DQP Level	Related DQP Data
BDQP	Create	New big data quality project	0	Metadata, Quality Requirements, …
	Re-use	An existing BDQP	All	
1	Add	Data sampling strategy	0	Sampling parameters
1	Add/update	Data profiling	1	Data profile (schema, statistical metric ratios scores)
2	Add/update	EQP (Predefined quality scenarios actions)	2	EQP parameters QR Proposals List
	Add	Qualitative QE (PCA, Feature Selection, etc.)		QLQE parameters (Attributes Sets)
	Update	QLQE attributes sets combination		(Combined Set)
3	Add/update	Mapping attributes and DQD's evaluation settings parameters	3	(DQES)
4	Update	Samples quantitative QE of DQD	4	QTQE results (DQES + Scores)
	Re-use/update	DQES Reused for QTQE of Pre-processed Samples (S')	7	(S' DQES + Scores)
5	Control	S DQD Scores vs Requirements S' DQD Scores vs Requirements	5 7	(Valid and Invalid Scores)
6	Add	Quality rules discovery from S DQES + Scores	6	(Quality Rules List)
7	Apply	Quality rules application by pre-processing Samples	7	Pre-processed Samples set S'
7	Validate	Analyze and check valid rules	7	(Valid and Invalid Quality Rules)
8	Optimize	Valid quality rules optimization	8	(QR optimized)
9	Apply	Big data pre-processing using optimized quality rules list	NA	New pre-processed Dataset DS'
10	Re-use/control/update	QTQE using DQES for DS' Samples, Score control	10	Quality report

(a) **Sampling:** Several sampling strategies can be applied to Big Data as surveyed in [54, 55]. In this work, the authors evaluated the effect of sampling methods on Big Data and concluded that the sampling of large datasets reduces the run-time and computational footprint of link prediction algorithms, maintaining an adequate prediction performance. In statistics, the Bootstrap sampling technique evaluates the sampling distribution of an estimator using sampling, which replaces the original samples. In the Big Data context, Bootstrap sampling has been studied in several works [56, 57]. In the proposed data quality evaluation scheme, it was decided to use the Bag of Little Bootstrap (BLB) [58]. This combines the results of bootstrapping multiple small subsets of a Big Data dataset. The BLB algorithm employs an original Big Dataset, which is used to generate small samples without replacements. For each generated sample, another set of samples is created by re-sampling with replacements.

(b) **Profiling:** The data profiling module performs the data quality screening based on statistics and information summary [59–61]. Since profiling is meant to discover data characteristics from data sources, it is considered as a data assessment process that provides a first summary of the data quality reported in its data profile.

Such information includes, for example, data format description, different attributes their types, values, and basic quality dimensions' evaluations, data constraints (if any), and data ranges (max and min, a set of specific values or subsets).

More precisely, the information about the data is presented in two types: technical and functional data. This information can be extracted from the data itself without any additional representation using metadata or any descriptive header file or by parsing the data using analysis tools. This task may become very costly in Big Data. Therefore, to avoid costs generated by the data size, the same sampling process (based on BLB) is used. Thus, the data is reduced to a representative population sample, in addition to the combination of profiling results. More precisely, a data profile in the proposed framework is represented as a data quality profile of the first level (**DQP1**), which is generated after the profiling phase. Moreover, data profiling provides some useful information that leads to significant data quality rules, usually named as data constraints. These rules are mostly equivalent to a structured-data schema, which is represented as technical and functional rules.

According to [61], there are many activities and techniques used to profile the data. These may range from online, incremental, and structural, to continuous profiling. Profiling tasks aim at discovering information about the data schema. Some data sources are already provided with their data profiles, sometimes with minimal information. In the following, some other techniques are introduced. These techniques can enrich and bring value-added information to a data profile:

- ***Data provenance inquiry:*** it tracks the data origin and provides information about data transformations, data copying, and its related data quality through the data lifecycle [62–64].
- ***Metadata:*** it provides descriptive and structural information about the data. Many data types, such as images, videos, and documents, use metadata to provide deep information about their contents. Metadata can be represented in many formats, including XML, or it can be extracted directly from the data itself without any additional representation.
- ***Data parsing (supervised/manual/automatic):*** data parsing is required since not all the data has a provenance or metadata that describes the data. The hardest way to gather extra information about the data is to parse it. Automatic parsing can be initially applied. Then, it is tuned and supervised manually by a data expert. This task may become very costly when Big Data is concerned, especially in the case of unstructured data. Consequently, a data profile is generated to represent only certain parts of the data that make sense. Therefore, multiple data profiles for multiple data partitions must be taken into consideration.
- ***Data profile:*** it is generated early in the Big Data Project as DQP Level 0 (Data profile in its early form) and upgraded as a data quality profile within the data preparation component as DQP Level 1. Then, it is updated and extended through all the components of the Big Data Quality Management Framework until it reaches a **DQP Level 2**. The **DQP Level 8** is the profile applied to the data in the pre-processing phase with its quality rules and related activities to output a pre-processed data conformed to the quality requirements.

Table 7 Example of exploratory quality profiling scenarios

#	Predefined scenario actions (PPAF)	Scenario actions PPA functions description	Execution order
1	*DeleteCols(**dqd,“>”,TL**)*	*Drop all columns or attributes with **dqd** ratio greater than **TL**: Tolerance Level*	1
2	*DeleteRows(**dqd,“>”,TL**)*	*Drop all Observations or rows with **dqd** ratio greater than **TL***	1

3	*DeleteRows(**dqd,“>”,TL**)*
	*DeleteCols(**dqd,“>”,TL**)*
		*Drop all Observations, then Attributes with **dqd** ratio greater than **TL***	1
			2

4	*DeleteRows(**dqd,“>”,TL**)*	*Drop all Observations or rows with **dqd** ratio greater than **TL***	1
	***newdqd** = Re-Evaluate ()*	*Recalculate the new **dqd** ratio after the row drop*	2
	*DeleteCols(**newdqd,“>”,TL**)*	*Drop all columns or attributes with **newdqd** ratio greater than **TL***	3
5	*DeleteCols(**dqd,“>”,TL**)*	*Drop all attributes with **dqd** ratio greater than **TL***	1
	***newdqd** = Re-Evaluate ()*	*Recalculate the **new dqd** ratio after the attributes drop*	2
	*DeleteRows(new**dqd,“>”,TL**)*	*Drop all observations with **newdqd** ratio greater than TL*	3

Exploratory quality profiling

Since a data-driven approach that uses a quantitative approach to quality dimensions’ evaluation from the data itself is followed, two evaluation steps are adopted: Quantitative Quality Evaluation based on user requirements and Exploratory Quality Profiling.

The exploratory quality profiling component is responsible for automatic data quality dimensions’ exploration without user interventions. The Quality Rules Proposals module, which produces a list of actions to elevate data quality, is based on some elementary DQDs that fit all varieties and data types.

A list of quality rules proposition, which is based on the quality evaluation of the most likely considered DQDs (e.g., completeness, accuracy, and uniqueness), is produced. This preliminary assessment is performed based on the data itself and using predefined scenarios. These scenarios are meant to increase data quality for some basic DQDs. In Fig. 7, the steps involved in the exploratory quality profiling for quality rules proposals generation are depicted. DQP1 is extended to DQP2, after adding the Data Quality Rules Proposal (***DQRP***), which is generated by the “quality rules proposals” process.

This module is part of the DQ profiling process, which varies the DQD tolerance levels from min to max scores and applies a systematic list of predefined quality rules. These predefined rules are a set of actions applied to the data when the measured DQD scores are not in the tolerance level defined by the min, max value scores. The actions vary from deleting only attributes, discarding only observations, or a combination of both. After these actions, a re-evaluation of the new DQD scores will lead to a quality rules proposal (DQRP) with known DQD target scores after performing an analysis. In Table 7, some examples of these predefined rules scenarios for the DQD completeness (***dqd = Comp***) with an execution priority for each set of grouped actions, are described. The DQD levels are set to vary from a 5% to 95% tolerance score with a granularity step of 5. They can be set differently according to the DQD choice and its sensitivity to the data model and domain. The selection of the best-proposed data quality rules is based on the KNN algorithm using Euclidean distance (Deng et al. 2016.; [65]. It gives the closest quality rules parameters that achieve (by default) high completeness with less data reduction. The process might be refined by specifying other quality parameters.

A list of quality rules proposal based on quality evaluation of the most likely considered DQD's (e.g., completeness, accuracy, and uniqueness), is produced. This preliminary assessment is based on the data itself using predefined scenarios. The quality rules are meant to increase data quality for some basic DQD's. In Fig. 8, the modules involved in the exploratory quality profiling for quality rules proposals generation, are illustrated.

Quality mapping and selection

The quality mapping and selection module of the BDQM framework is responsible for mapping data features or attributes to DQD's to target pre-required quality evaluation scores. It generates a Data Quality Evaluation Scheme (**DQES**) and then adds it (updates) to the DQP. The DQES contains the DQD's of the appropriate attributes to be evaluated using adequate metric formulas. The DQES, as a part of DQP, contains (for each of the selected data attributes) the following list, which is considered essential for the quantitative quality evaluation:

- The attributes: all or a selected list,
- The data quality dimensions (DQD's) to be evaluated for each selected attribute,
- Each DQD has a metric that returns the quality score, and
- The quality requirement scores for each DQD needed in the score's validation.

These requirements are general and target many global quality levels. The mapping component acts as a refinement of the global settings with precise qualities' goals. Therefore, a mapping must be performed between the data quality dimensions and targeted data features/attributes before proceeding with the quality assessment. Each DQD is measured for each attribute and sample. The mapping generates a ***DQES***, which contains Quality Evaluation Requests (**QER**) $\boldsymbol{Q_x}$. Each ***QER*** $\boldsymbol{Q_x}$ targets a data quality dimension (DQD) for an attribute, all attributes, or a set of selected attributes, where x is the number of requests.

(a) **Quality mapping:** Many approaches are available to accomplish an efficient mapping process. These include automatic, interactive, manual, and based on quality rules proposals techniques:

- ***Automatic****:* it completes the alignment and comparison of the data attributes (from DQP) with the data quality requirements (either per attribute type, or name). A set of DQDs is associated with each attribute for quality evaluation. It results in a set of associations to be executed and evaluated in the quality assessment component.
- ***Interactive****:* it relies on experts' involvement to refine, amend, or confirm the previous automated associations.
- ***Manual****:* it uses a similar but advanced dashboard to that illustrated in Error! Reference source not found. and a more detailed one in the attribute level.
- ***Quality rules proposals****:* the proposal list collected from the DQP2 is used to obtain an understanding of the impact of a DQD level and the data reduction

ratio. These quality insights help decide which DQD is best when compared to the quality requirements.

b) **Quality selection (of DQD, Metrics and Attributes):** It consists of a selection of an appropriate quality metric to evaluate data quality dimensions for an attribute of a Big Data sample set and returns a count of correct values, which comply with the metric formula. Each metric will be computed if the attribute values reflect the DQD constraints. For example, accuracy can be defined as a count of correct attributes in a certain range of values [v_1, v_2]. Similarly, it can be defined to satisfy a certain number of constraints related to the type of data such as zip code, email, social security number, dates, or addresses.

Let us define the tuple ***DQES (S, D, A, M)***. Most of the information is provided by the ***BDQP(DS,DS', Req) with Req*** = (***D***, ***L, A***) parameters. The profiling information is used to select the appropriate quality metrics $\boldsymbol{m_l}$ to evaluate the data quality dimensions $\boldsymbol{q_l}$ for an attribute $\boldsymbol{a_k}$ with a weight $\boldsymbol{w_j}$. In addition to the previous settings, let us consider the following:***S***: ***S***(***DS***,***N***, ***n, R***)→ $\boldsymbol{S_i}$ a sampling strategy

- Let us denote by ***M***, a set of quality metrics $\boldsymbol{M} = \{\boldsymbol{m}_1, .., \boldsymbol{m}_l, .., \boldsymbol{m}_d\}$ where $\boldsymbol{m_l}$ is a quality metric that measures and evaluates a DQD $\boldsymbol{q_l}$ for each value of an attribute $\boldsymbol{a_k}$ in the sample $\boldsymbol{s_i}$ and returns 1, if correct, and 0, if not. Each $\boldsymbol{m_l}$ metric will be computed if the value of the attribute reflects the $\boldsymbol{q_l}$ constraint. For example, the accuracy of an attribute is defined as a range of values between 0 and 100. Otherwise, it is incorrect. If the same DQD $\boldsymbol{q_l}$ is evaluated for a set of attributes, and if the weights are all equal, a simple mean is computed. The metric $\boldsymbol{m_l}$ will be evaluated to measure if each attribute has its $\boldsymbol{m_l}$ correct. This is performed for each instance (cell or row) of the sample $\boldsymbol{s_i}$.
- Let us denote by $\boldsymbol{M_l}^{(i)}, i = 1, \ldots, \boldsymbol{N}$, a metric total$\boldsymbol{m_l}$, which evaluates and counts the number of observations that satisfy this metric, for a DQD $\boldsymbol{q_l}$ of an attribute $\boldsymbol{a_k}$ of ***N*** samples from the dataset ***DS***. The proportion of observations under the adequacy rule is calculated by:

$$\boldsymbol{p_{k,i}} = \frac{\boldsymbol{M_l}^{(i)}(\boldsymbol{a_k})}{\boldsymbol{n}}, \boldsymbol{k} = 1, \ldots, \boldsymbol{R}$$

The proportion of observations under the adequacy rule in a sample $\boldsymbol{s_i}$ is given by:

$$\boldsymbol{p_i} = \sum_{\boldsymbol{k}=1}^{\boldsymbol{R}} \boldsymbol{p_{k,i}}$$

The total proportion of observations under the adequacy rule for all samples is given by:

$$\boldsymbol{M_l} = \frac{1}{\boldsymbol{N}} \sum_{\boldsymbol{i}=1}^{\boldsymbol{N}} \boldsymbol{p_i} = \frac{1}{\boldsymbol{N}} \sum_{\boldsymbol{i}=1}^{\boldsymbol{N}} \sum_{\boldsymbol{k}=1}^{\boldsymbol{R}} \frac{\boldsymbol{M_l}^{(i)}(\boldsymbol{a_k})}{\boldsymbol{n}}$$

where $\boldsymbol{M_l}$ characterizes the $\boldsymbol{q_l}$ mean score for the whole dataset.

- Let $\boldsymbol{Q_x}(\boldsymbol{a_k}, \boldsymbol{q_l}, \boldsymbol{m_l})$ represents a request for a quality evaluation, which results in the mean quality score for a DQD $\boldsymbol{q_l}$ for a measurable attribute $\boldsymbol{a_k}$ calculated by $\boldsymbol{M_l}$. The process by which Big Data samples are evaluated for a DQD $\boldsymbol{q_j}$ in a sample $\boldsymbol{s_i}$ for an attribute $\boldsymbol{a_k}$ with a metric $\boldsymbol{m_l}$, providing a $\boldsymbol{q_l s_i}$ score for each sample (described below in *Quantitative Quality Evaluation*). Then, a sample mean $\boldsymbol{q_l}$ is the final score for $\boldsymbol{a_k}$.
- Let us denote a process, which sorts and combines the requests of a quality evaluation (QER) by DQD or by an attribute, resulting in a re-arrangement of the $\boldsymbol{Q_x}(\boldsymbol{a_k}, \boldsymbol{q_l}, \boldsymbol{m_l})$ tuple into two types, depending on the evaluation selection group parameter:

1. Per DQD identified as $\boldsymbol{Q_x}(\boldsymbol{AList}(\boldsymbol{a_z}), \boldsymbol{q_l}, \boldsymbol{m_l})$ where $\boldsymbol{AList(a_z)}$ represents the attributes $\boldsymbol{a_z}$ (***z:1…R***) to be evaluated for the DQD $\boldsymbol{q_l}$.
2. Per attributes identified as $\boldsymbol{Q_x(a_k, DList(q_l, m_l))}$, where $\boldsymbol{DList(q_l, m_l)}$ represents the data quality dimensions $\boldsymbol{d_l}$(***l:1… d***) to be evaluated for the attribute $\boldsymbol{a_k}$.
   In some cases, the type of combination is automatically selected for a certain DQD, such as consistency, when all the attributes are constrained towards specific conditions. The combination is either based on attributes or DQD's, and the DQES will be constructed as follows:

DQES ($\boldsymbol{Q_x}(\boldsymbol{AList}(\boldsymbol{a_z}), \boldsymbol{q_l}, \boldsymbol{m_l})$,…,…) or.
DQES ($\boldsymbol{Q_x}(\boldsymbol{a_k}, \boldsymbol{DList}(\boldsymbol{q_l}, \boldsymbol{m_l}))$,…,…)

- The completion of the quality mapping process updates the DQP Level 2 with a ***DQES*** set as follows (Also illustrated in Error! Reference source not found.):

  DQES ($\boldsymbol{Q_x}(\boldsymbol{a_k}, \boldsymbol{q_l}, \boldsymbol{m_l})$***,…,…)***, where ***x*** ranges from **1** to a defined number of evaluation requests. Each $\boldsymbol{Q_x}$ element is a quality evaluation request of an attribute $\boldsymbol{a_k}$ for a quality dimension $\boldsymbol{q_l}$, with a DQD metric $\boldsymbol{m_l}$.

  The output of this phase generates a DQES score, which contains the mean score for each DQ dimension for one or many attributes. The mapping and selection data flow initiated using Big Data quality project (BDQP) settings is illustrated in Fig. 9. This is accomplished either using the same BDQP ***Req*** or defining more detailed and refined quality parameters and a sampling strategy. Two types of DQES can be yielded:
- Data Quality Dimension-wise evaluation of a list of attributes or
- Attribute-wise evaluation of many DQD's. As described before, the quality mapping and selection component generates a DQES evaluation scheme for the dataset, identifying which DQD and attributes tuples to evaluate using a specific quality metric. Therefore, a more detailed and refined set of parameters can also be set, as described in previous sections. In the following, the steps that construct the DQES in the mapping component are depicted:
- The ***QMS*** function extracts the ***Req*** parameters from ***BDQP*** as ***(D, L, A)***.
- A quality evaluation request (a_k, q_l, m_l), is generated from the ***(D, A)*** tuple.
- A list is constructed with these quality evaluation requests.

- A list sorting is performed either by DQD or by Attributes producing two types of lists:

   a. A combination of requests per DQD generates quality requests for a set of attributes $(AList(a_z), q_l, m_l)$.
   b. A combination of requests per attribute generates quality requests for a set of DQD's $(a_k, DList(q_l, m_l))$.

- A DQES is returned based on the evaluation selection group parameter (per DQD, per attribute).

Quantitative quality evaluation

The Authors in [66], addressed how to evaluate a set of DQDs over a set of attributes. According to this study, the evaluation of Big Data quality is applied and iterated to many samples. The aggregation and combination of DQD's scores are performed after each iteration. The evaluation scores are added to the DQES, which results in updating the DQP. We proposed an algorithm, which computes the quality scores for a dataset based on a certain quality mapping and quality metrics.

This algorithm is based on quality metrics evaluation using scores after collecting and validating the scores with quality requirements and generating quality rules from these scores [66, 67]. There are rules related to each pre-processing activity, such as data cleaning rules, which eliminate data, and data enrichment, which replaces or adds data. Other activities, such as data reduction, reduce the data size by decreasing the number of features or attributes that have certain characteristics such as low variance, and highly correlated features.

In this phase, all the information collected from previous components (profiling, mapping, DQES) is included in the data quality profile level 3. The important elements are the set of samples and the data quality evaluation scheme, which are executed on each sample to evaluate its quality attributes for a specific DQD.

DQP Level 3 provides all the information needed about the settings represented by the DQES to proceed with the quality evaluation. The DQES contains the following:

- The selected DQDs and their related metrics.
- The selected attributes with the DQD to be evaluated.
- The DQD selection, which is based on the Big Data quality requirements expressed early when initiating a Big Data Quality Project.
- Attributes selection is set in the quality selection mapping component (3).

The quantitative quality evaluation methodology is described as follows:

a. The selected DQD quality metrics will measure and evaluate the DQD for each attribute observation in each sample from the sample set. For each attribute observation, it returns a value 1, if correct, or 0, if incorrect.
b. Each metric will be computed if all the sample observations attribute values reflect the constraints. For example, the metric accuracy of an attribute defines that a range

of values between 20 and 70 is valid. Otherwise, it is invalid. The count of correct values out of the total sample observations is the DQD ratio represented by a percentage (%). This is performed for all selected attributes and their selected DQDs.

c. The sample mean from all samples for each evaluated DQD represents a Data Quality Score (DQS) estimation $\left(\overline{DQS}\right)$ of a data quality dimension of the data source.
d. **DQP Level 4**: an update to the DQP level 3 includes a data quality evaluation scheme (DQES) with the quality scores per DQD and per attribute (***DQES + Scores***).
e. In summary, the quantitative quality evaluation starts with sampling, DQD's and DQDs metrics selection, mapping with data attributes, quality measurements, and the sample mean DQD's ratios.

Let us denote by $\boldsymbol{Q_x}$ ***Score*** (quality score), the evaluation results of each quality evaluation request $\boldsymbol{Q_x}$ in the ***DQES***. Two types of **DQES,** depending on the evaluation type, which means two kind of results scores organized per DQD of all attributes or per attribute for all DQD's, can be identified:

$\boldsymbol{Q_x\left(AList(a_z), q_l, m_l\right) \to Q_x\, ScoreList\left(AList(a_z, Score), q_l, m_l\right)}$ or.

$\boldsymbol{Q_x\left(a_z, DList(q_l, m_l)\right)} \to \mathrm{Q_xScoreList}\left(\boldsymbol{a_z, DList\left(q_l, m_l, Score\right)}\right)$

where $\boldsymbol{z} = 1, \ldots, \boldsymbol{r}, \boldsymbol{r}$ is the number of selected attributes, and $\boldsymbol{l} = 1, \ldots, \boldsymbol{d}$, $\boldsymbol{d}$ is the number of selected DQD's.

The quality evaluation generates quality scores $\boldsymbol{Q_x}$ ***Score***. A quality scoring model is used to assess these results. It is provided in the form of quality requirements to comprehend the resulted scores, which are expressed as quality acceptance level percentages. These quality requirements might be a set of values, or an interval in which values are accepted or rejected, or a single score ratio percentage. The analysis of these scores against quality requirements leads to the discovery and generation of quality rules for attributes violating the quality requirements.

The quantitative quality evaluation process follows the steps described below for the case of the evaluation of a DQD's list among several attributes $\left(\boldsymbol{Q_x\left(a_z, DList(q_l, m_l)\right)}\right)$:

1) ***N*** samples (of size ***n***) are generated from the dataset ***DS*** using a **BLB-based** bootstrap sampling approach.
2) For each sample $\boldsymbol{s_i}$ generated in step 1, and
3) For each $\boldsymbol{a_z}$ $(\boldsymbol{z} = 1, \ldots, \boldsymbol{r})$ selected attribute in ***DQES*** in step 1, evaluate all the DQD's in the ***DList*** using their related metrics to obtain ***Q$_x$ScoreList*** $\left(\boldsymbol{a_z, DList\left(q_l, m_l, Score\right), s_i}\right)$ for each sample $\boldsymbol{s_i}$.
4) For all the samples scores, evaluate the sample mean of all ***N*** samples for each attribute $\boldsymbol{a_z}$ related to the $\boldsymbol{q_l}$ evaluation scores, as $\overline{\overline{\boldsymbol{q}}_{\boldsymbol{zl}}}$.
5) For the dataset ***DS***, evaluate the quality score mean $\overline{\boldsymbol{q}}_{\boldsymbol{l}}$ for each DQD for all attributes $\boldsymbol{a_z}$, as follows:

$$\overline{\boldsymbol{q}}_{\boldsymbol{l}} = 1/\boldsymbol{r}\sum\nolimits_{\boldsymbol{z}=1}^{\boldsymbol{r}} \overline{\boldsymbol{q}}_{\boldsymbol{zl}}$$

The illustration in Fig. 10 shows that the $\boldsymbol{q_{zl}s_iScore}$ is the evaluation of DQD $\boldsymbol{q_l}$ for the sample $\boldsymbol{s_i}$ for an attribute $\boldsymbol{a_z}$ with a metric $\boldsymbol{m_l}\overline{\boldsymbol{q}}_{\boldsymbol{zl}}$ represents the quality score sample mean for the attributes $\boldsymbol{a_z}$.

Quality control

The quality control is initiated when the quality evaluation results are available and reported in the **DQES** in **DQP Level 4**. During quality control, all the quality scores with the quality requirements of the Big Data project are checked. If any detected anomalies or a non-conformance are found, the quality control component forwards a **DQP Level 5** to the data quality rules discovery component.

At this point, various cases are highlighted. An iteration process is performed until the required quality levels are satisfied, or the experts decide to stop the quality evaluation process and re-evaluate their requirements. At each phase, there is a kind of quality control, even if it is not explicitly specified, within each quality process.

The quality control acts in the following cases:

Case 1: This case applies when the quality is estimated, and no rules are yet included in the **DQP Level 4** (the DQP is considered as a report, since the data quality is still inspected, and only reports are generated with no actions yet to be performed).

a. In the case of accepted quality scores, no quality actions need to be applied to data. The **DQP Level 4** remains unchanged and acts as a full data quality report, which is updated with positive validation of the data per quality requirement. However, it might include some simple pre-processing such as attribute selection and filtering. According to the data analytics requirements and expected results planned in the Big Data project, more specific data pre-processing actions are performed but not related to quality in this case.
b. In the case when quality scores are not accepted, the **DQP Level 4 DQES** scores are analyzed, and the DQP is updated with a quality error report about the related DQD scores and their data attributes. **DQP Level 5** is created, and it will be analyzed by the quality rules discovery component for the pre-processing activities to be executed on the data.

Case 2: In the presence of a **DQP Level 6** that contains a quality evaluation request of the pre-processed samples with discovered quality rules, the following situations may occur:

a. When the quality control checks that the **DQP Level 6** rules are valid and satisfy the quality requirements, the **DQP Level 6** is updated to **DQP Level 7** and confirmed as the final data quality profile, which will be applied to the data in the pre-processing phase. **DQP Level 7** is considered as important if it contains validated quality rules.
b. When the quality control is not totally or partially satisfied, the **DQP Level 6** is sent back for an adaptation of the quality selection and mapping component with valid and invalid quality rules, quality scores, and error reports. These reports highlight with an unacceptable score interval the quality rules that have not satisfied the quality requirements. The quality selection and mapping component provide automatic or manual analysis and assessment of the unsatisfied quality rules concerning their targeted DQD's, attributes, and quality requirements. An adaptation of quality requirements is needed to re-validate these rules. Finally, the user experts have the

final word to continue or break the process and proceed to the pre-processing phase with the valid rules. As part of the framework reuse specification, the invalid rules are kept within the DQP for future re-evaluation.

Case 3: The control component will always proceed based on the quality scores and quality requirements for both input and pre-processed data. Continuous control and monitoring are responsible for initiating DQP updates and adaptation if the quality requirements are relaxed.

Quality rules, discovery, validation, optimization, and execution

In [67] work, it was reported that if the DQD scores do not conform to the quality requirements, then failed scores are used to discover data quality rules. When executed on data, these rules enhance its quality. They are based on known pre-processing activities such as data cleansing. Each activity has a set of functions targeting different types of data in order to increase its DQD ratio and the whole Data Quality (of the Data source or the Dataset(s)).

When Quality Rules (**QR)** are applied to a sample set **S**, a pre-processed sample set ***S'*** is generated. A quality evaluation process is invoked on ***S'***, generating DQD scores for ***S'***. Thus, a score comparison between ***S*** and ***S'*** is conducted to filter only qualified and valid rules with a higher percentage of success among data. Then, an optimization scheme is applied to the list of valid quality rules before their application on production data. The predefined optimization schemes vary from (1) rules priority to (2) rules redundancy, (3) rules removal, (4) rules grouping per attribute, or (5) per DQD's, or (6) per duplicate rules.

a) **Quality rules discovery:** The discovery is based on the **DQP Level 5** from the quality control component. An analysis of the quality scores is initiated, and an error report is extracted. If the DQD scores do not conform to the quality requirements, then failed scores are used to discover data quality rules. When executed on data, these rules enhance its quality. They are based on known pre-processing activities such as data cleansing. Error! Reference source not found. illustrates the several modules of the discovery component from DQES DQDs scores analysis versus requirements, attributes pre-processing activities combination for each targeted DQD, and the rules generation.

   For example, an attribute having a 50% score of missing data is not accepted for a required score of 20% or less. This initiates the generation of a quality rule, which consists of a data cleansing activity for observations that do not satisfy the quality requirements. The data cleansing or data enrichment activity is selected from the Big Data quality profile repository. The quality rule will target all the related attributes marked for pre-processing to reduce the 50% to 20% for the DQD completeness. Moreover, in the case of completeness, not only cleansing can be applied to missing values, but many alternatives are available for pre-processing activities. These activities are related to completeness such as missing values replacement activity with many functions for several replacements' methods like the mean, mode, and the median.

The pre-processing activities are provided by the repository to achieve the required data quality. Many possibilities for pre-processing activities selection are available:

- *Automatic*, by discovering and suggesting a set of activities or DQ rules.
- *Predefined*, by selecting ready-to-use quality rules proposals from the exploratory quality profiling component, predefined pre-processing activity functions from the repository, indexed by DQDs.
- *Manual,* giving the expert the ability to query the exploratory quality profiling results for the best rules, achieving the required quality using KNN-based filtering.

b) **Quality rules validation:** The generated quality rules from the discovery components are set in the DQP **Level 6.** As illustrated in Error! Reference source not found., the rules validation component process starts when the DQR list is applied to the sample set ***S***, resulting in a pre-processed sample set ***S′***, which is generated by the related pre-processing activities. Then, a quality evaluation process is invoked on ***S′***, generating DQD scores for ***S′***. Thus, a score comparison between ***S*** and ***S′*** is conducted to filter only qualified and valid rules with a higher percentage of success among data. After analyzing this score, two sets of rules are identified: successful and failed rules.

c) **Quality rules optimization:** After the set of discovered valid quality rules is selected, an optimization process is activated to reorganize and filter the rules. This is due to the nature of the evaluation parameters set in the mapping component and the refinement of the quality requirement. These choices with the rule's validation process will produce a list of individual quality rules that, if applied as generated, might have the following consequences:

- Redundant rules.
- Ineffective rules due to the order of execution.
- Multiple rules, which target the same DQD with the same requirements.
- Multiple rules, which target the same attributes for the same DQD and requirements.
- Rules, which drop attributes or rows, must be applied first or have a higher priority to avoid applying rules on data items that are meant to be dropped (Table 8).

The quality rules optimization component applies an optimization scheme to the list of valid quality rules before their application to production data in the pre-processing phase. The predefined optimization schemes vary according to the following, as illustrated in Error! Reference source not found.:

- Rules execution priority per attribute or DQD, per pre-processing activity, or pre-processing function.
- Rules redundancy removal per attributes or DQDs.
- Rules grouping, combination, per activity, per attribute, per DQD's, or duplicates.
- For invalid rules, the component consists of several actions, including rules removal or rules adaptation from previously generated proposals in the exploratory quality profiling component for the same targeted tuple (attributes, DQDs).

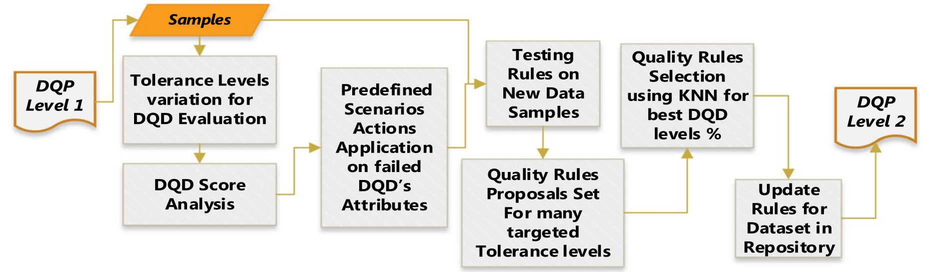

Samples
DQP Level 1
Tolerance Levels variation for DQD Evaluation
DQD Score Analysis
Predefined Scenarios Actions Application on failed DQD's Attributes
Testing Rules on New Data Samples
Quality Rules Proposals Set For many targeted Tolerance levels
Quality Rules Selection using KNN for best DQD levels %
Update Rules for Dataset in Repository
DQP Level 2

Fig. 8 Quality rules proposals with exploratory quality profiling

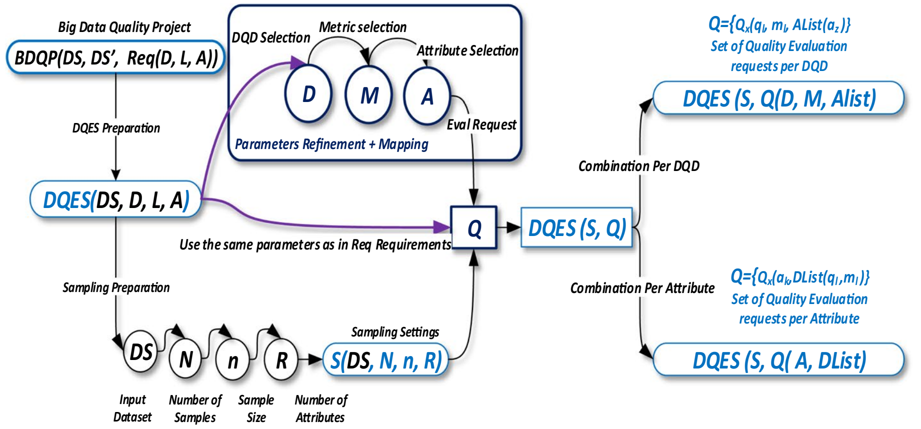


Fig. 9 DQES parameters settings

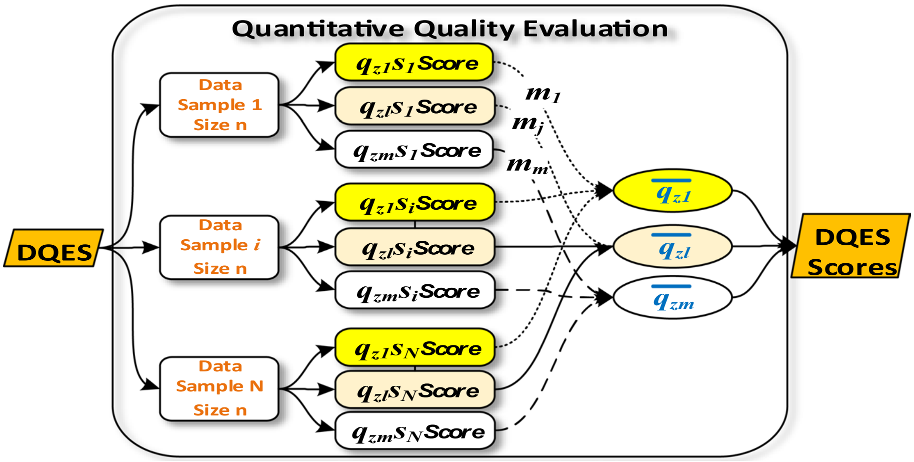


Fig. 10 Big data sampling and quantitative quality evaluation

(d) **Quality rules optimization:** The Quality Rules execution consists of pre-processing data using the DQP, which embeds the data quality rules that enhance the quality to reach the agreed requirements. As part of the monitoring module, a sampling set from the pre-processed data is used to re-assess the quality and detect eventual failures.

Table 8 Quality rules optimization schemes

Rules' optimization methods	Per			
	Attribute	DQD	PPA	PPAF
Execution priority	●	●	●	●
Redundancy removal	●	●		
Rules grouping and combination	●	●	●	●
Invalid rules removal and adaptation	●	●	●	●

PP Pre-processing, *PPA* Pre-processing activity, *PPAF* Pre-Processing activity function

Quality monitoring

Quality Monitoring is a continuous quality control process, which relies on the DQP. The purpose of monitoring is to validate the DQP across all the Big Data lifecycle processes. The QP repository is updated during and after the complete lifecycle as well as after the user's feedback data, quality requirements, and mapping.

As illustrated in Fig. 11, the monitoring process takes a scheduled snapshot of the pre-processed Big Data all along the BDQMF for the BDQ project. This data snapshot is a set of samples that have their quality evaluated in the BDQMF component (4). Then, quality control is conducted on the quality scores, and an update is performed to the DQP. The quality report may highlight the quality failure and its ratio evolution through multiple sampling snapshots of data.

The monitoring process strengthens and enforces the quality across the Big Data value chain using the BDQM framework while reusing the data quality profile information. For each quality monitoring iteration on the datasets from the data source, quality reports are added to the data quality profile, updating it to a **DQP Level 10**.

Data processing, analytics, and visualization

This process involves the application of algorithms or methodologies, which extract insights from the ready-to-use data, with enhanced quality. Then, the value of processed data is projected visually as a dashboard and graphically enhanced charts for the decision-makers to act economically. Big Data visualization approaches are of high importance for the final exploitation of the data.

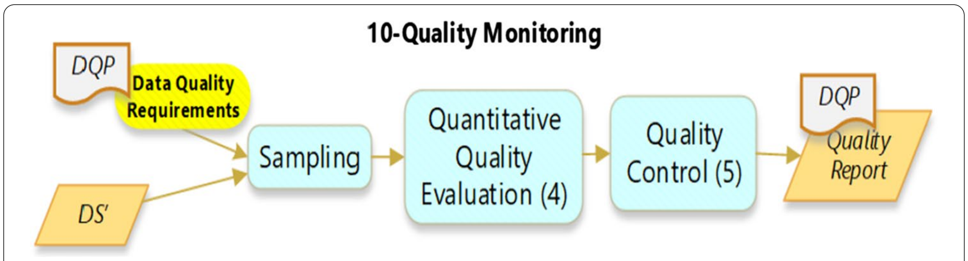


Fig. 11 Quality monitoring component

Implementations: Dataflow and quality processes development

In this section, we overview the dataflow across the various processes of the framework, we also highlight the implemented quality management processes along with the supporting application interfaces developed to support main processes. Finally, we describe the ongoing processes' implementations and evaluations.

Framework dataflow

In Fig. 12, we illustrate the whole process flow of the framework, from the inception of the quality project in its specification and requirements to the quality monitoring phase. As an ongoing process, monitoring is a part of the quality enforcement loop and may trigger other processes that handle several quality profile operations like DQP adaptation, upgrade, or reuse.

In Table 9, we enumerate and detail the multiple processes and their interactions within the BDQM Framework components including their inputs and outputs after executing related activities with the quality profile (DQP), as detailed in the previous section.

Quality management processes' implementation

In this section, we describe the implementation of our framework's important components, processes, and their contributions towards the quality management of Big Data across its lifecycle.

Core processes implementation

As depicted above, core framework processes have been implemented and evaluated, in the following, we describe how these components have been implemented and evaluated.

Quality profiling: one of the central components of our framework is the data quality profile (DQP). Initially, the DQP implements a simple data profile of a Big Data set as an XML file (DQP Sample illustrated in Fig. 13).

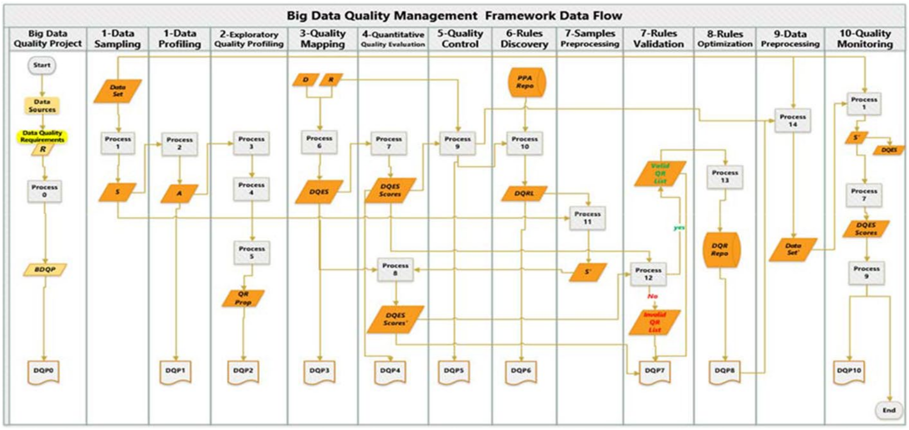


Fig. 12 Big data quality management framework data flow

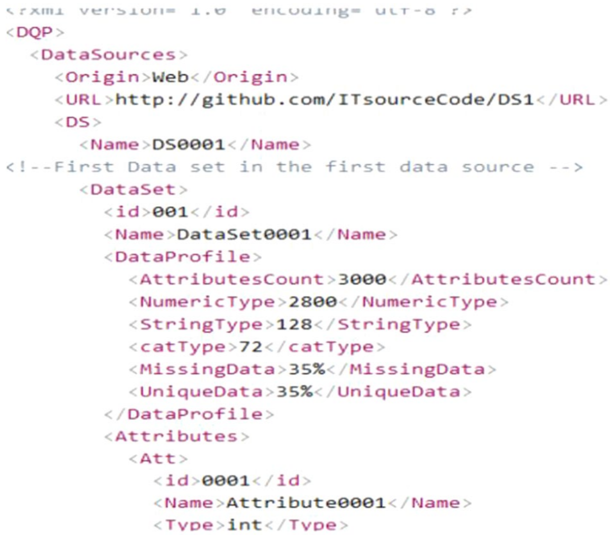
Table 9 Data quality profile levels dataflow

Proc #	BDQMF comp #	Description	Input	Output
Start	BDQP	Big Data quality project (BDQR) creation with quality requirements R, data sources (DS)	R, DS	DQP 0 (BDQP(DS, R))
1	1	Sampling strategy parameters (sample size, number)	BDQP	DQP 0 Samples set S
2	1	Data profiling	S	DQP 1 (Data Profile)
3	2	EQP: Quality rules Proposals scenarios (Sc) based	Sc, S	DQP 2 (QR Proposals)
4	2	QQE: Best ranked attributes selection lists	S	DQP 2 (Attributes Sets)
5	2	QQE: Combination of lists of best attributes	S, Sets	DQP 2 (Combined Set)
6	3	Data quality evaluation scheme specification	R, D	DQP 3 (DQES)
7	4	Quantitative quality evaluation of dataset samples	S, DQES	DQP 4 (DQES + Scores)
8	4	Quantitative quality evaluation of preprocessed samples	S', DQES	DQP 7 (S'DQD Scores)
9	5	Control of DQES DQD scores	R, DQES + Scores	DQP 5 (DQD OK, Not)
10	6	Quality Rules' discovery based on DQES scores	DQES, PPA_QPREPO	DQP 6 (Quality Rules List)
11	7	Preprocessing samples using discovered QR	QR List, S	Preprocessed Samples set S'
12	7	Quality Rules' validation	S, S' DQES + Scores	DQP 7 (Valid, Not Valid Quality Rules)
13	8	Quality Rule's optimization	DQP 7	DQP 8 (QR optimized)
14	9	Big data preprocessing	Dataset DS	Dataset DS'
End/Loop	10	Quality monitoring	DS' samples	DQP 10 Quality Report

DS the dataset, *R* Requirements, *DQPx* Data quality Profile Level

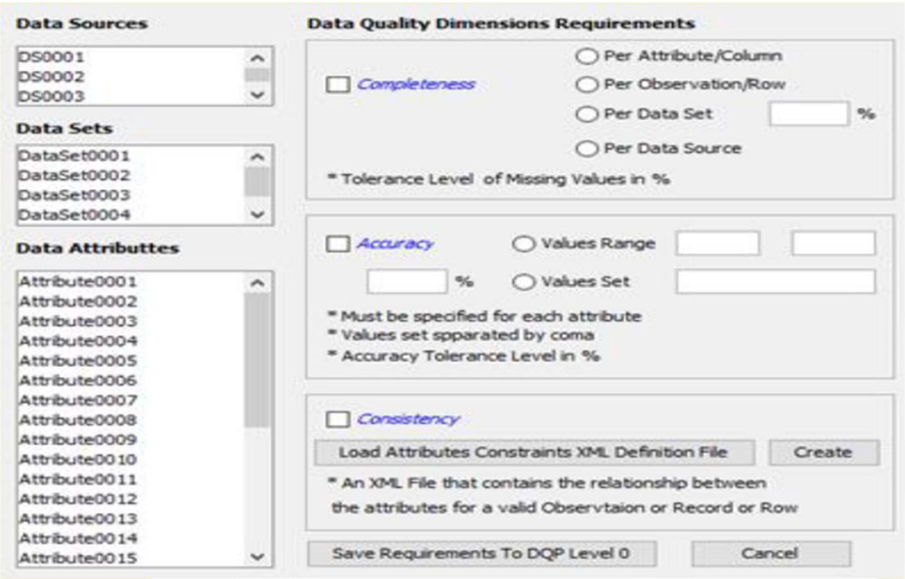
After traversing several framework component's processes, it is updated to a data quality profile. The data quality evaluation process is one of the activities that updates the DQP with quality scores that are later used to discover data quality rules. These rules, when applied to the original data, will ensure an output data set with higher quality. The DQP is finally executed by the pre-processing component. Through the end of the lifecycle, the DQP contains all pieces of information such as data quality rules that target a set of data sources with multiple datasets, data attributes and data quality dimensions such as accuracy, and pre-processing activities like data cleansing, data integration, and data normalization. The Data Quality Profile (DQP) contains all the information about the Data, its Quality, the User Quality Requirements, DQD's, Quality Levels, Attributes, the Data Quality Evaluation Scheme (DQES), Quality Scores, and the Data Quality Rules. The DQP is stored in the DQP repository, which contains the following modules, and performs many tasks related to DQP. In the following, the DQP lifecycle and its repository are described.

Quality requirement dashboard: developed as a web-based application as shown in Fig. 14 below to capture user's requirements and other quality information. Such



```
<?xml version="1.0" encoding="utf-8"?>
<DQP>
  <DataSources>
    <Origin>Web</Origin>
    <URL>http://github.com/ITsourceCode/DS1</URL>
    <DS>
      <Name>DS0001</Name>
<!--First Data set in the first data source -->
      <DataSet>
        <id>001</id>
        <Name>DataSet0001</Name>
        <DataProfile>
          <AttributesCount>3000</AttributesCount>
          <NumericType>2800</NumericType>
          <StringType>128</StringType>
          <catType>72</catType>
          <MissingData>35%</MissingData>
          <UniqueData>35%</UniqueData>
        </DataProfile>
        <Attributes>
          <Att>
            <id>0001</id>
            <Name>Attribute0001</Name>
            <Type>int</Type>
```

Fig. 13 Example of data quality profile



Data Sources

DS0001
DS0002
DS0003

Data Sets

DataSet0001
DataSet0002
DataSet0003
DataSet0004

Data Attributtes

Attribute0001
Attribute0002
Attribute0003
Attribute0004
Attribute0005
Attribute0006
Attribute0007
Attribute0008
Attribute0009
Attribute0010
Attribute0011
Attribute0012
Attribute0013
Attribute0014
Attribute0015

Data Quality Dimensions Requirements

□ Completeness
○ Per Attribute/Column
○ Per Observation/Row
○ Per Data Set %
○ Per Data Source
* Tolerance Level of Missing Values in %

□ Accuracy
○ Values Range
%
○ Values Set
* Must be specified for each attribute
* Values set spparated by coma
* Accuracy Tolerance Level in %

□ Consistency
Load Attributes Constraints XML Definition File
Create
* An XML File that contains the relationship between the attributes for a valid Observtaion or Record or Row

Save Requirements To DQP Level 0
Cancel

Fig. 14 Quality requirements dashboard

requirements include for instance data quality dimension requirements specification. This application can be extended with extra information about data sources such as attributes and their types. The user is guided through the interface to specify the right attributes' values and also given the option to upload an XML file containing the relationship between attributes. The recorded requirements are finally saved to a data quality profile level 0 which will be used in the next stage of the quality management process.

Data preparation and sampling: The framework operations start when the quality project's minimal specifications are set. It initiates and provides a data quality summary named data quality profile (DQP) by running an exploratory quality profiling assessment on data samples (using BLB sampling algorithm). This DQP is projected to be the core component of the framework and every update and every result regarding the quality is noted/recorded. The DQP is stored in a quality repository and registered in the Big Data's provenance to keep track of data changes due to quality enhancements.

Data quality mapping and rule discovery components: data quality mapping alleviates and adds more data quality control to the whole data quality assessment process. The implemented mapping links and categorizes all the quality project required elements, from Big Data quality characteristics, pre-processing activities, and their related techniques functions, to data quality rules, dimensions, and their metrics. The Data Quality Rules' discovery from evaluation results implementation reveals the required actions and transformations that when applied on the data set will accomplish the targeted quality level. These rules are the main ingredients of pre-processing activities. The role of a DQ rule is to undertake the sources of bad quality by defining a list of actions related to each quality score. The DQ rules are the results of systematic and planned data quality assessment analysis.

Quality profile repository (QPREPO): Finally, our framework implements the *QPREPO* to manage the data quality profiles for different data types and domains and to adapt or optimize existing profiles. This repository manages the data quality dimensions with their related metrics, and the pre-processing activities, and their activity functions. A QPREPO entry is implemented for each Big Data quality project with the related DQP containing information's about each dataset, data source, data domain, and data user. This information is essential for DQP reuse, adaptation, and enhancement for the same or different data sources.

Implemented approaches for quality assessment.

The framework uses various approaches for quality assessment: (1) Exploratory Quality Profiling; (2) a Quantitative Quality Assessment approach using DQD metrics; and it's anticipated to add a new component for (3) a Qualitative quality assessment.

(1) Exploratory Quality Profiling implements an automatic quality evaluation that is done systematically on all data attributes for basic DQDs. The resulted in calculated scores are used to generate quality rules for each quality tolerance ratio variation. These rules are then applied to other data samples and the quality is reassessed. An analysis of the results provides an interactive quality-based rules search using several ranking algorithms (maximization, minimization, applying weight).

(2) The Quantitative Quality Assessment implements a quick data quality evaluation strategy supported through sampling and profiling processes for Big Data. The evaluation is conducted by measuring the data quality dimensions (DQDs) for attributes using specific metrics to calculate a quality score.

(3) The Qualitative Quality Assessment approach implements a deep quality assessment to discover hidden quality aspects and their impact on the Big Data Lifecycle outputs. These quality aspects must be quantified into scores and mapped with related attributes and DQD's. This quantification is achieved by applying several feature selection strategies and algorithms to data samples. These qualitative

insights are combined with those obtained before the quantitative quality evaluation early in the Quality management process.

Framework development, deployment, and evaluation

Development, deployment, and evaluation of our BDQMF framework follow a systematic modular approach where various components of the framework are developed and tested independently then integrated with the other components to compose the integrated solution. Most of the components are implemented in R and |Python using SparkR and PySpark libraries respectively. The supporting files like the DQP, DQES, and configuration files are written in XML and JSON formats. Big Data quality project requests and constraints including the data sources and the quality expectation are implemented within the solution where more than one module might be involved. The BDQMF components are deployed following Apache Hadoop and Spark ecosystem architecture.

The BDQMF deployed modules implementation description and developed APIs are listed in the following:

- **Quality setting mapper (QSP):** it implements an interface for automatic selection and mapping of DQD's and dataset attributes from the initial DQP.
- **Quality settings parser (QSP):** responsible for parsing and loading parameters to the execution environment from DQP settings to data files. It is also used to extract quality rules and scores from the DQES in the DQP.
- **Data loader (DL):** implements filtering, selecting, and loading all types of data files required by the BDQMF including datasets from data sources into the Spark environment (e.g. DataFrames, tables), it will be used by various processes or it will persist in the database for further reuse. For data selection the uses SQL to retrieve only attributes being set in the DQP settings.
- **Data samples generator (DSG):** it generates data samples from multiple data sources.
- **Quality inspector and profiler (QIP):** it is responsible for all qualitative and quantitative quality evaluations among data samples for all the BDQMF lifecycle phases. The inspector assesses all the default and required DQD's, and all quality evaluations are set into the DQES within the DQP file.
- **Preprocessing activities and functions execution engine (PPAF-E**): all the repository preprocessing activities along with their related functions are implemented as APIs in python and R. When requested this library will load the necessary methods and execute them within the preprocessing activities for rules validation and rules execution in phase 9.
- **Quality rules manager (QRM):** it is one of the important modules of the framework. It implements and deliver the following features:

  - Analyzes Quality results
  - Discovers and generates Quality rules proposals.
  - Quality rules validation among requirements settings.

- o Quality rules refinement and optimizations
- o Quality rules ACID operations in the DQP files and the repository.

- **Quality monitor (QM)**: it is responsible for monitoring, triggering, and reporting any quality change all over the Big Data lifecycle to assure the efficiency of quality improvement of the discovered data quality rules.
- **BDQMF-Repo:** is the repository where all the quality-related files, settings, requirements, results are stored. The repo is using HBase or Mongo DB to fulfill requirements of the Big Data ecosystem environments and scalability for intensive data updates.

Conclusion

Big data quality has attracted the attention of researchers regarding Big Data as it is considered the key differentiator, which leads to high-quality insights and data-driven decisions. In this paper, a Big Data Quality Management Framework for addressing end-to-end Quality in the Big Data lifecycle was proposed. The framework is based on a Data Quality Profile, which is augmented with valuable information while traveling across different stages of the framework, starting from Big Data project parameters, quality requirements, quality profiling, and quality rules proposals. The exploratory quality profiling feature, which extracts quality information from the data, helped in building a robust DQP with a quality rules proposal and a step over for the configuration of the data quality evaluation scheme. Moreover, the extracted quality rules proposals are of high benefit for the quality dimensions mapping and attribute selection component. This fact supports the users with quality data indicators characterized by their profile.

The framework dataflow shows that any Big Data set quality is evaluated through the exploratory quality profiling component and the quality rules extraction and validation towards an improvement in its quality. It is of great importance to ensure the right selection of a combination of targeted DQD levels, observations (rows), and attributes (columns) for efficient quality results, while not sacrificing vital data because of considering only one DQD. The resulted quality profile based on the quality assessment results confirms that the contained quality information significantly improves the quality of Big Data.

In future work, we plan to extend the quantitative quality profiling with qualitative evaluation. We also plan to extend the framework to cope with unstructured Big Data quality assessment.

Acknowledgements
Not applicable.

Authors' contributions

IT conceived the main conceptual ideas related to Big data quality framework and proof outline. He designed the framework and their main modules, he also worked on the implementation and validation of some of the framework's components. MAS supervised the study and was in charge of direction and planning, he also contributed to couple of sections including the introduction, abstract, the framework and the implementation and conclusion section. CB contributed to data preparation sampling and profiling, he also reviewed and validated all formulations and statistical modeling included in this work. RD contributed in the review and discussion of the core contributions and their validation. All authors read and approved the final manuscript.

Authors' information

Dr. Ikbal Taleb is currently an Assistant Professor, College of Technological Information, Zayed University, Abu Dhabi, U.A.E. He got his Ph.D. in information and systems engineering from Concordia University in 2019, and MSc. in Software Engineering from the University of Montreal, Canada in 2006. His research interests include data and Big data quality, quality profiling, quality assessment, cloud computing, web services, and mobile web services.

Prof. M. Adel Serhani is currently a Professor, and Assistant Dean for Research and Graduate Studies College of Information Technology, U.A.E University, Al Ain, U.A.E. He is also an Adjunct faculty in CIISE, Concordia University, Canada. He holds a Ph.D. in Computer Engineering from Concordia University in 2006, and MSc. in Software Engineering from University of Montreal, Canada in 2002. His research interests include: Cloud for data intensive e-health applications, and services; SLA enforcement in Cloud Data centers, and Big data value chain, Cloud federation and monitoring, Non-invasive Smart health monitoring; management of communities of Web services; and Web services applications and security. He has a large experience earned throughout his involvement and management of different R&D projects. He served on several organizing and Technical Program Committees and he was the program Co-Chair of International Conference in Web Services (ICWS'2020), Co-chair of the IEEE conference on Innovations in Information Technology (IIT´13), Chair of IEEE Workshop on Web service (IWCMC´13), Chair of IEEE workshop on Web, Mobile, and Cloud Services (IWCMC´12), and Co-chair of International Workshop on Wireless Sensor Networks and their Applications (NDT´12). He has published around 130 refereed publications including conferences, journals, a book, and book chapters.

Dr. Chafik Bouhaddioui is an Associate Professor of Statistics in the College of Business and Economics at UAE University. He got his Ph.D. from University of Montreal in Canada. He worked as lecturer at Concordia University for 4 years. He has a rich experience in applied statistics in finance in private and public sectors. He worked as assistant researcher in Finance Ministry in Canada. He worked as Senior Analyst in National Bank of Canada and developed statistical methods used in stock market forecasting. He joined in 2004 a team of researchers in finance group at CIRANO in Canada to develop statistical tools and modules in finance and risk analysis. He published several papers in well-known journals in multivariate time series analysis and their applications in economics and finance. His area of research is diversified and includes modeling and prediction in multivariate time series, causality and independence tests, biostatistics, and Big Data.

Prof. Rachida Dssouli is a full professor and Director of Concordia Institute for Information Systems Engineering, Faculty of Engineering and Computer Science, Concordia University. Dr. Dssouli received a Master (1978), Diplome d'études Approfondies (1979), Doctorat de 3eme Cycle in Networking (1981) from Université Paul Sabatier, Toulouse, France. She earned her PhD degree in Computer Science (1987) from Université de Montréal, Canada. Her research interests are in Communication Software Engineering a sub discipline of Software Engineering. Her contributions are in Testing based on Formal Methods, Requirements Engineering, Systems Engineering, Telecommunication Service Engineering and Quality of Service. She published more than 200 papers in journals and referred conferences in her area of research. She supervised/ co-supervised more than 50 graduate students among them 20 PhD students. Dr. Dssouli is the founding Director of Concordia Institute for Information and Systems Engineering (CIISE) June 2002. The Institute hosts now more than 550 graduate students and 20 faculty members, 4 master programs, and a PhD program.

Funding

This work is supported by fund #12R005 from ZCHS at UAE University.

Availability of data and materials

Data used in this work is available with the first author and can be provided up on request. The data includes sampling data, pre-processed data, etc.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]College of Technological Innovation, Zayed University, P.O. Box 144534, Abu Dhabi, United Arab Emirates. [2]College of Information Technology, UAE University, P.O. Box 15551, Al Ain, United Arab Emirates. [3]Department of Statistics, College of Business and Economics, UAE University, P.O. Box 15551, Al Ain, United Arab Emirates. [4]Concordia Institute for Information Systems Engineering, Concordia University, Montreal, QC H4B 1R6, Canada.

Received: 6 February 2021 Accepted: 15 May 2021
Published online: 29 May 2021

References

1. Chen M, Mao S, Liu Y. Big data: A survey. Mobile Netw Appl. 2014;19:171–209. https://doi.org/10.1007/s11036-013-0489-0.
2. Chiang F, Miller RJ. Discovering data quality rules. Proceed VLDB Endowment. 2008;1:1166–77.
3. Yeh, P.Z., Puri, C.A., 2010. An Efficient and Robust Approach for Discovering Data Quality Rules, in: 2010 22nd IEEE International Conference on Tools with Artificial Intelligence (ICTAI). Presented at the 2010 22nd IEEE International Conference on Tools with Artificial Intelligence (ICTAI), pp. 248–255. https://doi.org/10.1109/ICTAI.2010.43
4. Ciancarini, P., Poggi, F., Russo, D., 2016. Big Data Quality: A Roadmap for Open Data, in: 2016 IEEE Second International Conference on Big Data Computing Service and Applications (BigDataService). Presented at the 2016 IEEE Second International Conference on Big Data Computing Service and Applications (BigDataService), pp. 210–215. https://doi.org/10.1109/BigDataService.2016.37
5. Firmani D, Mecella M, Scannapieco M, Batini C. On the meaningfulness of "big data quality" (Invited Paper). Data Sci Eng. 2016;1:6–20. https://doi.org/10.1007/s41019-015-0004-7.
6. Rivas, B., Merino, J., Serrano, M., Caballero, I., Piattini, M., 2015. I8K|DQ-BigData: I8K Architecture Extension for Data Quality in Big Data, in: Advances in Conceptual Modeling, Lecture Notes in Computer Science. Presented at the International Conference on Conceptual Modeling, Springer, Cham, pp. 164–172. https://doi.org/10.1007/978-3-319-25747-1_17
7. Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A.H., 2011. Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute 1–137.
8. Chen CP, Zhang C-Y. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. Inf Sci. 2014;275:314–47.
9. Hashem IAT, Yaqoob I, Anuar NB, Mokhtar S, Gani A, Ullah Khan S. The rise of "big data" on cloud computing: Review and open research issues. Inf Syst. 2015;47:98–115. https://doi.org/10.1016/j.is.2014.07.006.
10. Hu H, Wen Y, Chua T-S, Li X. Toward scalable systems for big data analytics: a technology tutorial. IEEE Access. 2014;2:652–87. https://doi.org/10.1109/ACCESS.2014.2332453.
11. Wielki J. The Opportunities and Challenges Connected with Implementation of the Big Data Concept. In: Mach-Król M, Olszak CM, Pełech-Pilichowski T, editors. Advances in ICT for Business. Springer International Publishing: Industry and Public Sector, Studies in Computational Intelligence; 2015. p. 171–89.
12. Ali-ud-din Khan, M., Uddin, M.F., Gupta, N., 2014. Seven V's of Big Data understanding Big Data to extract value, in: American Society for Engineering Education (ASEE Zone 1), 2014 Zone 1 Conference of The. Presented at the American Society for Engineering Education (ASEE Zone 1), 2014 Zone 1 Conference of the, pp. 1–5. https://doi.org/10.1109/ASEEZone1.2014.6820689
13. Kepner, J., Gadepally, V., Michaleas, P., Schear, N., Varia, M., Yerukhimovich, A., Cunningham, R.K., 2014. Computing on masked data: a high performance method for improving big data veracity, in: 2014 IEEE High Performance Extreme Computing Conference (HPEC). Presented at the 2014 IEEE High Performance Extreme Computing Conference (HPEC), pp. 1–6. https://doi.org/10.1109/HPEC.2014.7040946
14. Saha, B., Srivastava, D., 2014. Data quality: The other face of Big Data, in: 2014 IEEE 30th International Conference on Data Engineering (ICDE). Presented at the 2014 IEEE 30th International Conference on Data Engineering (ICDE), pp. 1294–1297. https://doi.org/10.1109/ICDE.2014.6816764
15. Gandomi A, Haider M. Beyond the hype: Big data concepts, methods, and analytics. Int J Inf Manage. 2015;35:137–44.
16. Pääkkönen P, Pakkala D. Reference architecture and classification of technologies, products and services for big data systems. Big Data Research. 2015;2:166–86. https://doi.org/10.1016/j.bdr.2015.01.001.
17. Oliveira, P., Rodrigues, F., Henriques, P.R., 2005. A Formal Definition of Data Quality Problems., in: IQ.
18. Maier, M., Serebrenik, A., Vanderfeesten, I.T.P., 2013. Towards a Big Data Reference Architecture. University of Eindhoven.
19. Caballero, I., Piattini, M., 2003. CALDEA: a data quality model based on maturity levels, in: Third International Conference on Quality Software, 2003. Proceedings. Presented at the Third International Conference on Quality Software, 2003. Proceedings, pp. 380–387. https://doi.org/10.1109/QSIC.2003.1319125
20. Sidi, F., Shariat Panahy, P.H., Affendey, L.S., Jabar, M.A., Ibrahim, H., Mustapha, A., 2012. Data quality: A survey of data quality dimensions, in: 2012 International Conference on Information Retrieval Knowledge Management (CAMP). Presented at the 2012 International Conference on Information Retrieval Knowledge Management (CAMP), pp. 300–304. https://doi.org/10.1109/InfRKM.2012.6204995
21. Chen, M., Song, M., Han, J., Haihong, E., 2012. Survey on data quality, in: 2012 World Congress on Information and Communication Technologies (WICT). Presented at the 2012 World Congress on Information and Communication Technologies (WICT), pp. 1009–1013. https://doi.org/10.1109/WICT.2012.6409222
22. Batini C, Cappiello C, Francalanci C, Maurino A. Methodologies for data quality assessment and improvement. ACM Comput Surv. 2009;41:1–52. https://doi.org/10.1145/1541880.1541883.
23. Glowalla, P., Balazy, P., Basten, D., Sunyaev, A., 2014. Process-Driven Data Quality Management–An Application of the Combined Conceptual Life Cycle Model, in: 2014 47th Hawaii International Conference on System Sciences (HICSS). Presented at the 2014 47th Hawaii International Conference on System Sciences (HICSS), pp. 4700–4709. https://doi.org/10.1109/HICSS.2014.575
24. Wand Y, Wang RY. Anchoring data quality dimensions in ontological foundations. Commun ACM. 1996;39:86–95. https://doi.org/10.1145/240455.240479.
25. Wang, R.Y., Strong, D.M., 1996. Beyond accuracy: What data quality means to data consumers. Journal of management information systems 5–33.
26. Cappiello, C., Caro, A., Rodriguez, A., Caballero, I., 2013. An Approach To Design Business Processes Addressing Data Quality Issues.
27. Hazen BT, Boone CA, Ezell JD, Jones-Farmer LA. Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications. Int J Prod Econ. 2014;154:72–80. https://doi.org/10.1016/j.ijpe.2014.04.018.

28. Caballero, I., Verbo, E., Calero, C., Piattini, M., 2007. A Data Quality Measurement Information Model Based On ISO/IEC 15939., in: ICIQ. pp. 393–408.
29. Juddoo, S., 2015. Overview of data quality challenges in the context of Big Data, in: 2015 International Conference on Computing, Communication and Security (ICCCS). Presented at the 2015 International Conference on Computing, Communication and Security (ICCCS), pp. 1–9. https://doi.org/10.1109/CCCS.2015.7374131
30. Woodall P, Borek A, Parlikad AK. Data quality assessment: The hybrid approach. Inf Manage. 2013;50:369–82. https://doi.org/10.1016/j.im.2013.05.009.
31. Goasdoué, V., Nugier, S., Duquennoy, D., Laboisse, B., 2007. An Evaluation Framework For Data Quality Tools., in: ICIQ. pp. 280–294.
32. Caballero, I., Serrano, M., Piattini, M., 2014. A Data Quality in Use Model for Big Data, in: Indulska, M., Purao, S. (Eds.), Advances in Conceptual Modeling, Lecture Notes in Computer Science. Springer International Publishing, pp. 65–74. https://doi.org/10.1007/978-3-319-12256-4_7
33. Cai L, Zhu Y. The challenges of data quality and data quality assessment in the big data era. Data Sci J. 2015. https://doi.org/10.5334/dsj-2015-002.
34. Philip Woodall, A.B., 2014. An Investigation of How Data Quality is Affected by Dataset Size in the Context of Big Data Analytics.
35. Laranjeiro, N., Soydemir, S.N., Bernardino, J., 2015. A Survey on Data Quality: Classifying Poor Data, in: 2015 IEEE 21st Pacific Rim International Symposium on Dependable Computing (PRDC). Presented at the 2015 IEEE 21st Pacific Rim International Symposium on Dependable Computing (PRDC), pp. 179–188. https://doi.org/10.1109/PRDC.2015.41
36. Liu, J., Li, J., Li, W., Wu, J., 2016. Rethinking big data: A review on the data quality and usage issues. ISPRS Journal of Photogrammetry and Remote Sensing, Theme issue "State-of-the-art in photogrammetry, remote sensing and spatial information science" 115, 134–142. https://doi.org/10.1016/j.isprsjprs.2015.11.006
37. Rao, D., Gudivada, V.N., Raghavan, V.V., 2015. Data quality issues in big data, in: 2015 IEEE International Conference on Big Data (Big Data). Presented at the 2015 IEEE International Conference on Big Data (Big Data), pp. 2654–2660. https://doi.org/10.1109/BigData.2015.7364065
38. Zhou, H., Lou, J.G., Zhang, H., Lin, H., Lin, H., Qin, T., 2015. An Empirical Study on Quality Issues of Production Big Data Platform, in: 2015 IEEE/ACM 37th IEEE International Conference on Software Engineering (ICSE). Presented at the 2015 IEEE/ACM 37th IEEE International Conference on Software Engineering (ICSE), pp. 17–26. https://doi.org/10.1109/ICSE.2015.130
39. Becker, D., King, T.D., McMullen, B., 2015. Big data, big data quality problem, in: 2015 IEEE International Conference on Big Data (Big Data). Presented at the 2015 IEEE International Conference on Big Data (Big Data), IEEE, Santa Clara, CA, USA, pp. 2644–2653. https://doi.org/10.1109/BigData.2015.7364064
40. Maślankowski, J., 2014. Data Quality Issues Concerning Statistical Data Gathering Supported by Big Data Technology, in: Kozielski, S., Mrozek, D., Kasprowski, P., Małysiak-Mrozek, B., Kostrzewa, D. (Eds.), Beyond Databases, Architectures, and Structures, Communications in Computer and Information Science. Springer International Publishing, pp. 92–101. https://doi.org/10.1007/978-3-319-06932-6_10
41. Fürber, C., Hepp, M., 2011. Towards a Vocabulary for Data Quality Management in Semantic Web Architectures, in: Proceedings of the 1st International Workshop on Linked Web Data Management, LWDM '11. ACM, New York, NY, USA, pp. 1–8. https://doi.org/10.1145/1966901.1966903
42. Corrales DC, Corrales JC, Ledezma A. How to address the data quality issues in regression models: a guided process for data cleaning. Symmetry. 2018;10:99.
43. Fan, W., 2008. Dependencies revisited for improving data quality, in: Proceedings of the Twenty-Seventh ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems. ACM, pp. 159–170.
44. Kläs, M., Putz, W., Lutz, T., 2016. Quality Evaluation for Big Data: A Scalable Assessment Approach and First Evaluation Results, in: 2016 Joint Conference of the International Workshop on Software Measurement and the International Conference on Software Process and Product Measurement (IWSM-MENSURA). Presented at the 2016 Joint Conference of the International Workshop on Software Measurement and the International Conference on Software Process and Product Measurement (IWSM-MENSURA), pp. 115–124. https://doi.org/10.1109/IWSM-Mensura.2016.026
45. Rahm E, Do HH. Data cleaning: Problems and current approaches. IEEE Data Eng Bull. 2000;23:3–13.
46. Dallachiesa, M., Ebaid, A., Eldawy, A., Elmagarmid, A., Ilyas, I.F., Ouzzani, M., Tang, N., 2013. NADEEF: A Commodity Data Cleaning System, in: Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data, SIGMOD '13. ACM, New York, NY, USA, pp. 541–552. https://doi.org/10.1145/2463676.2465327
47. Ebaid A, Elmagarmid A, Ilyas IF, Ouzzani M, Quiane-Ruiz J-A, Tang N, Yin S. NADEEF: A generalized data cleaning system. Proceed VLDB Endowment. 2013;6:1218–21.
48. Elmagarmid, A., Ilyas, I.F., Ouzzani, M., Quiané-Ruiz, J.-A., Tang, N., Yin, S., 2014. NADEEF/ER: generic and interactive entity resolution. ACM Press, pp. 1071–1074. https://doi.org/10.1145/2588555.2594511
49. Tang N. Big Data Cleaning. In: Chen L, Jia Y, Sellis T, Liu G, editors. Web Technologies and Applications. Lecture Notes in Computer Science: Springer International Publishing; 2014. p. 13–24.
50. Ge M, Dohnal V. Quality management in big data informatics. 2018;5:19. https://doi.org/10.3390/informatics5020019.
51. Jimenez-Marquez JL, Gonzalez-Carrasco I, Lopez-Cuadrado JL, Ruiz-Mezcua B. Towards a big data framework for analyzing social media content. Int J Inf Manage. 2019;44:1–12. https://doi.org/10.1016/j.ijinfomgt.2018.09.003.
52. Siddiqa A, Hashem IAT, Yaqoob I, Marjani M, Shamshirband S, Gani A, Nasaruddin F. A survey of big data management: Taxonomy and state-of-the-art. J Netw Comput Appl. 2016;71:151–66. https://doi.org/10.1016/j.jnca.2016.04.008.
53. Taleb, I., Dssouli, R., Serhani, M.A., 2015. Big Data Pre-processing: A Quality Framework, in: 2015 IEEE International Congress on Big Data (BigData Congress). Presented at the 2015 IEEE International Congress on Big Data (BigData Congress), pp. 191–198. https://doi.org/10.1109/BigDataCongress.2015.35

54. Cormode, G., Duffield, N., 2014. Sampling for Big Data: A Tutorial, in: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14. ACM, New York, NY, USA, pp. 1975–1975. https://doi.org/10.1145/2623330.2630811
55. Gadepally, V., Herr, T., Johnson, L., Milechin, L., Milosavljevic, M., Miller, B.A., 2015. Sampling operations on big data, in: 2015 49th Asilomar Conference on Signals, Systems and Computers. Presented at the 2015 49th Asilomar Conference on Signals, Systems and Computers, pp. 1515–1519. https://doi.org/10.1109/ACSSC.2015.7421398
56. Liang F, Kim J, Song Q. A bootstrap metropolis-hastings algorithm for bayesian analysis of big data. Technometrics. 2016. https://doi.org/10.1080/00401706.2016.1142905.
57. Satyanarayana, A., 2014. Intelligent sampling for big data using bootstrap sampling and chebyshev inequality, in: 2014 IEEE 27th Canadian Conference on Electrical and Computer Engineering (CCECE). Presented at the 2014 IEEE 27th Canadian Conference on Electrical and Computer Engineering (CCECE), IEEE, Toronto, ON, Canada, pp. 1–6. https://doi.org/10.1109/CCECE.2014.6901029
58. Kleiner, A., Talwalkar, A., Sarkar, P., Jordan, M., 2012. The big data bootstrap. arXiv preprint
59. Dai, W., Wardlaw, I., Cui, Y., Mehdi, K., Li, Y., Long, J., 2016. Data Profiling Technology of Data Governance Regarding Big Data: Review and Rethinking, in: Latifi, S. (Ed.), Information Technolog: New Generations. Springer International Publishing, Cham, pp. 439–450. https://doi.org/10.1007/978-3-319-32467-8_39
60. Loshin, D., 2010. Rapid Data Quality Assessment Using Data Profiling 15.
61. Naumann F. Data profiling revisited. ACM. SIGMOD Record. 2014;42:40–9.
62. Buneman, P., Davidson, S.B., 2010. Data provenance–the foundation of data quality.
63. Glavic, B., 2014. Big Data Provenance: Challenges and Implications for Benchmarking, in: Specifying Big Data Benchmarks. Springer, pp. 72–80.
64. Wang, J., Crawl, D., Purawat, S., Nguyen, M., Altintas, I., 2015. Big data provenance: Challenges, state of the art and opportunities, in: 2015 IEEE International Conference on Big Data (Big Data). Presented at the 2015 IEEE International Conference on Big Data (Big Data), pp. 2509–2516. https://doi.org/10.1109/BigData.2015.7364047
65. Hwang W-J, Wen K-W. Fast kNN classification algorithm based on partial distance search. Electron Lett. 1998;34:2062–3.
66. Taleb, I., Kassabi, H.T.E., Serhani, M.A., Dssouli, R., Bouhaddioui, C., 2016. Big Data Quality: A Quality Dimensions Evaluation, in: 2016 Intl IEEE Conferences on Ubiquitous Intelligence Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld). Presented at the 2016 Intl IEEE Conferences on Ubiquitous Intelligence Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld), pp. 759–765. https://doi.org/10.1109/UIC-ATC-ScalCom-CBDCom-IoP-SmartWorld.2016.0122
67. Taleb, I., Serhani, M.A., 2017. Big Data Pre-Processing: Closing the Data Quality Enforcement Loop, in: 2017 IEEE International Congress on Big Data (BigData Congress). Presented at the 2017 IEEE International Congress on Big Data (BigData Congress), pp. 498–501. https://doi.org/10.1109/BigDataCongress.2017.73
68. Deng, Z., Zhu, X., Cheng, D., Zong, M., Zhang, S., n.d. Efficient kNN classification algorithm for big data. Neurocomputing. https://doi.org/10.1016/j.neucom.2015.08.112
69. Firmani, D., Mecella, M., Scannapieco, M., Batini, C., 2015. On the Meaningfulness of "Big Data Quality" (Invited Paper), in: Data Science and Engineering. Springer Berlin Heidelberg, pp. 1–15. https://doi.org/10.1007/s41019-015-0004-7
70. Lee YW. Crafting rules: context-reflective data quality problem solving. J Manag Inf Syst. 2003;20:93–119.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.