Quality Evaluation for Big Data

A Scalable Assessment Approach and First Evaluation Results

Michael Kläs, Wolfgang Putz
Fraunhofer Institute for Experimental Software Engineering IESE
Kaiserslautern, Germany
{michael.klaes, wolfgang.putz}@iese.fraunhofer.de

Tobias Lutz
NTT DATA Deutschland GmbH
Ettlingen, Germany
tobias.lutz@nttdata.com

***Abstract*—High-quality data is a prerequisite for most types of analysis provided by software systems. However, since data quality does not come for free, it has to be assessed and managed continuously. The increasing quantity, diversity, and velocity that characterize big data today make these tasks even more challenging. We identified challenges that are specific for big data quality assessments with particular emphasis on their usage in smart ecosystems and make a proposal for a scalable cross-organizational approach that addresses these challenges. We developed an initial prototype to investigate scalability in a multi-node test environment using big data technologies. Based on the observed horizontal scalability behavior, there is an indication that the proposed approach also allows dealing with increasing volumes of heterogeneous data.**

Keywords—big data quality assessment; quality measurement; velocity; volume; variety; SQA4BD; QUAMOCO; smart ecosystems; SPARK; HADOOP

I. INTRODUCTION

Data is a central element of any software system. It is thus not surprising that current software quality standards do not only demand certain qualities from the software system itself [12] but also from the data managed by the system [13].

However, although data quality has been a research topic for several decades and although the introduction of business intelligence solutions has raised awareness of it not only on the operative level but also among company decision makers, there are estimates that companies still lose up to 25 percent of their operative gains due to data quality issues [4].

Today, the problem is exacerbated since the availability of data is increasing and companies are making first attempts to leverage the potential promised under the label of *big data*. Where means do exist for managing and assessing the quality of "small" data [14] [13] [2], they are not directly transferable to big data, which is not only larger in terms of quantity but also more diverse and volatile. Although a recent study considers data quality as a key enabler for emerging *big data ecosystems*, the same study states that it is still largely unclear how to assess and manage data quality in these settings [23].

Therefore, we see the need for an approach that facilitates *data quality assessments* in the context of big data. The key challenges are motivated by the properties of the assessed object, i.e., the 3Vs – *volume, variety, and velocity* – that make data into big data [20], rather than by the assumption that other qualities might be relevant for big data compared to "small" data. Additionally, big data is increasingly used in cross-organizational contexts, where data is considered as a *good* provided by a company with certain qualities and is further processed or used by others. In such ecosystems, as developed, for example, in the PRO-OPT project [27] for the automotive domain, limitations regarding data access and transmission have to be considered not only during data usage but also during data assessment.

The main contributions of this paper are:

- An approach for cross-organizational and scalable quality assessments for big data (SQA^{4}BD), which addresses challenges we identified [19] by (1) dealing with various types of data in a unified entity model using quality factors, (2) decoupling measurement and evaluation in the assessment, (3) making use of a horizontal scaling architecture, and (4) enabling incremental updates of assessments using delta measurement
- Preliminary evaluation results on our work by investigating the performance effects of horizontal scaling based on an initial prototype

The remaining paper is structured as follows: We introduce a generalized conceptual model for data quality assessments and discuss the implications of big data on data quality assessments (Section 2). Next, we review related work in the context of data quality assessments and highlight observed gaps (Section 3). Then we introduce SQA^{4}BD, our approach for scalable cross-organizational quality assessments for big data (Section 4). The scalability of the approach is evaluated based on an initial prototypical implementation (Section 5). Finally, we conclude the paper with a summary and outlook on future work (Section 6).

II. BACKGROUND

A. Conceptual Framework

This section presents a conceptual model for data quality assessment, which is summarized and illustrated in Fig. 1. We introduce basic concepts and provide a terminology that will be used in the following sections.

Quality is an abstract concept, and its definition changes dependent on the actual point of view: On the one hand, it can be understood as satisfying explicit as well as implicit needs of

the user, and on the other hand as complying with predetermined specifications [16]. On the ISO level, there are currently two standards ISO/TS 8000-1: 2011 and ISO/IEC 25012: 2008, which define data quality as follows: the "degree to which data meets user requirements" [14] and the "degree to which the characteristics of data satisfy stated and implied needs when used under specified conditions" [13].

This means that data quality has to be seen in strong connection to the projected use of data and can only be assessed in this context, but not as an isolated concept. Hence, the starting point for any quality assessment are the *Quality Needs*, which, inter alia, depend on the users of the data and their information needs (e.g., analysis goals and requirements).

Based on the quality needs, the abstract concept of data quality can be refined by a set of *Quality Aspects* with different degrees of importance. ISO/IEC 25012 defines data quality aspects (in the standard called Data Quality Characteristics) as "category of data quality attributes that bears on data quality" and provides in its data quality model definitions for fifteen quality aspects: accuracy, completeness, consistency, credibility, currentness, accessibility, compliance, confidentiality, efficiency, precision, traceability, understandability, availability, portability, and recoverability [13].

In order to support holistic quality assessment, rules for the aggregation of the resulting partial assessments on the quality aspect level should also be available in addition to the refinement of quality aspects. Depending on the underlying quality needs, these could be weighting and aggregation rules for quality aspects.

Quality aspects are quantified by *Measures*. According to ISO/TS 8000-1, these are "variables to which a value is assigned as the result of measurement" [14]. They quantify specific characteristics of a data construct. An example measure of the quality aspect *completeness* could be the relative number of missing values in a table column.

The concrete specification of the measure depends on the nature of the assessed data and the available metadata ("data that describe other data" [13]). Examining a quality aspect like *precision*, for instance, requires different kinds of quantification depending on the type of data (image data, text, or floating point numbers). To what extent and in which form a quality aspect can be quantified and checked also depends on the available metadata. For example, in the absence of metadata on the data type of a measured characteristic, it could be unclear which mathematical operations can be applied. If metadata concerning the format of date information is not available (e.g., YYYY-MM-DD), it cannot be checked for consistency.

Measures may also be determined by means of different *Instruments*. An instrument realizes the means of collecting a measure. These may be manual means – if, for example, a person completes a questionnaire – or automatic means – if a tool analyzes the data. Possible tools include a script in a statistics application such as R [10], an SQL script directly executing in a database, such as MySQL [25], or a specialized data quality tool such as Talend Open Studio [30]. Examples of manual means are online surveys such as LimeSurvey [21] or classical paper questionnaires.

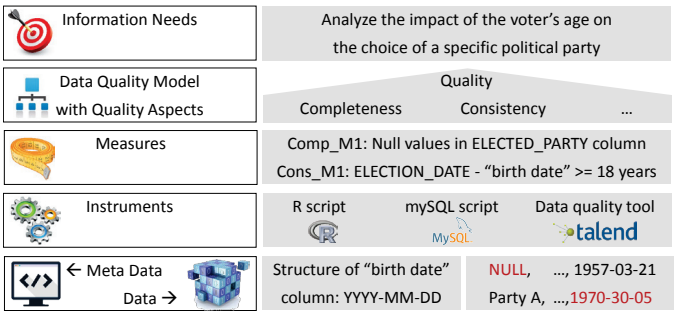

Fig. 1. Conceptual components of a data quality assessment

B. Challenges

In the literature, several characterizations of big data can be found, which differ in the number and denomination of Vs, but at least commonly include the following three "classical" properties: size (*volume*), heterogeneity (*variety*), and speed (*velocity*) [20]. Added Vs, such *veracity* (how meaningful are the data for the intended use and how reliable are the conclusions made?) and *value* (of which value are the analysis results?) represent on the other hand no intrinsic properties of data that would help to characterize the specific nature of big data but represent quality needs relevant for all kind of data.

Next, we will thus outline the effects of volume, variety, and velocity on data quality assessment and the resulting challenges. Moreover, reference will be made to the conceptual components of data quality assessment introduced above.

Volume: The most obvious feature of big data is the volume of the data. It reaches a level that can no longer be handled effectively with classical databases and analysis tools on a single computer. Dealing with such quantities of data is increasingly made possible by new technologies, which enable effective distribution of data across a network of computers (horizontal scaling) and hide the complexity of distributed data management from the user.

From a quality assessment perspective, this means that existing quality metrics have to be operationalized in a way that will allow their application to large distributed datasets. In most cases, this is not true for *existing instruments* because they need the data locally on the computer to analyze them (such as in classical R implementations) or they only be used with standard databases. Generally speaking, it can be deduced that the instruments implementing quality measures have to run performance-critical computations at the location where the data are stored and then aggregate the intermediate results into an overall result.

For a rapid preliminary analysis, an alternative could be to calculate the measures on a representative subset of the data. In such cases, the assessment results would have to be marked with an explicit level of uncertainty.

Velocity: Another important characteristic of big data is the speed with which new data is generated and has to be processed. In contrast to data analysis on the basis of traditional data warehouse solutions, such as those operated in many companies,

where large parts of datasets are extracted and analyzed in batch mode, big data applications often require real-time analysis (e.g., for preventive maintenance of production equipment) or they use data that are subject to change much more frequently or on a larger scale.

The high volatility of the data results in new challenges for data quality assessment. Instead of being re-assessed every few months, data quality must now be re-evaluated several times a day or even in near real-time. However, frequent re-assessments require an (almost) fully automatic measurement process with automated instruments. Using expert interviews as instruments for determining data quality as proposed by numerous existing assessment methods does not appear to be a practicable solution.

In addition, frequent re-evaluations combined with the relatively large volume of data induce potential performance problems in big data applications. In particular, the conventional approach of measuring the entire dataset in a batch process for each assessment is considered inappropriate.

Variety: A third important characteristic of big data is its heterogeneity. Data is no longer provided in a centralized and structured relational database for analysis purposes, but rather data from different data sources are integrated and analyzed jointly. These data sources are modeled differently, are provided in different formats, and are accessible via different technologies. In particular, semi-restructured data (such as websites) and even unstructured data (such as images, video, and text) are increasingly being used for analyses as well.

Existing quality models and measures still focus primarily on structured data, but measures that are suitable for structured data are not directly applicable to semi-structured or unstructured data. For example, the *precision* of an image cannot be quantified using the same measure used for the precision of a numeric value. Consequently, in many cases existing quality aspects must be backed by new measures to be determined for semi-structured or unstructured data.

Cross-organizational data usage can be considered a specific aspect of variety in big data applications. The constantly growing number of available data sources in the public space due to Open Data initiatives, as well as the ongoing development of ecosystems with cross-organizational use of data such as in PRO-OPT [27] and the Industrial Data Space [11], makes it impractical for a user to assess the quality of all the data sources planned for use in a specific application. Rather, it would be desirable for data to also include preliminary information on its quality as part of its descriptive metadata. This is even more important when data is not available for direct access and the terms and conditions of data usage have to be negotiated by the partners of the ecosystem.

At a basic level, such a quality assessment could be provided by a standardized quality model (incl. appropriate measures) and the obtained quality assessment results could be made available as part of the metadata. Ideally, mechanisms would be provided that allow potential data users to adapt the standardized quality model to their specific information needs in advance (e.g., to determine relevant quality aspects and their relative importance) and to get an individual quality assessment without accessing the data.

III. Related Work

In this section, the related work in the context of data quality assessments is reviewed with a special focus on its application on big data. However, such work is currently limited as observed, for instance, by Clarke, who focused his literature searches on sources likely to be read by researchers and practitioners working with big data. He concluded that there are *"multiple vacuous statements along the lines of 'data quality can be a problem', but very few sources that actually address the key questions of what the level of data quality is in 'big data', how it can be assessed"* [29].

We addressed the topic with a *systematic mapping study* [26] on data quality assessments in general, in which we identified 44 data quality models [18]. To gain insights into whether and how the models fit the big data challenges, a description scheme was developed and applied to the identified models.

In addition to more general aspects such as the application domain for which the model was developed or which quality aspects were considered, the scheme comprises aspects that allow inferring whether the model addresses big data applications. Regarding the big data challenges volume, variety, and velocity, the models were characterized, among others, by the attributes *object data type*, *object volume*, *object heterogeneity*, *object volatility*, *purpose*, *measures*, and *instruments*. Object data type characterizes whether the model is applicable for structured, semi-structured, or unstructured data. Object volume and object heterogeneity describe whether high data volume, respectively data heterogeneity, are explicit aspects of the model. Object volatility covers the volatility aspect. Purpose answers the question for which kind of usage the approach was developed (specification, measuring / monitoring, assessment), and measures and instruments describe whether concrete measures or instruments are part of the model. The complete description scheme as well as detailed results can be found in separate report [18].

To sum up the results, it can be stated that:

- Around three-quarters of the models do not explicitly define for which object data type they are applicable and only 9% explicitly address semi-structured objects. Unstructured objects are addressed by 4% of the models.
- Object volume and object heterogeneity are considered only in 14%, respectively 5% of the models.
- Regarding the purpose of the models, they cover specification (39%), measuring / monitoring (32%), and assessment (27%). In most cases, models supporting assessment also support measuring, and models supporting measuring also cover specification, which means that nearly all the models allow concretizing data quality by quality aspects.
- The definition of measures focuses on specific quality aspects and it can be observed that mainly frequently named aspects such as completeness, consistency, or free-of-error are also frequently concretized.
- Because half of the considered models propose no measures, there is no possibility for these models to give evidence on

instruments. In about one quarter of the models, however, measures can be determined automatically and in 14% semi-automatically.

- Even if data volume [9] [8], and data heterogeneity [24] [3] [7] [28] [6] [1], are addressed in some models, none of the models explicitly addresses the scalability challenge. This also means that there are no approaches that define measures and instruments with the intention of scaling to big distributed or volatile data.

Ongoing work explicitly addressing big data quality, such as the Big Data Quality Framework (BDQF) proposed by the United Nations Economic Commission for Europe [31] and the quality model proposed by Chai and Zhu [5] focuses on specifying models with quality aspects particularly relevant for big data rather than on how quality assessments can be applied in the context of big data. Thus, these activities are orthogonal to the work presented in this paper, which approaches the challenges for quality assessments caused by big data.

IV. The SQA⁴BD Approach

This section provides first an overview on the SQA⁴BD approach, then, details are given on the three major stages of a quality assessment, (1) providing a base model for data quality, (2) analyzing the data, and (3) evaluating quality.

A. Overview

Considering the assessment of big data in a cross-organizational scenario, three major roles can be identified: a kind of *authority* defining which aspects of quality could be relevant and how they are quantified for different kinds of data types, the *data provider*, who is willing to share certain data under conditions to be negotiated, and the potential *data consumer*, who is interested in making use of certain data offered by the provider depending on whether the data help to satisfy the consumer's information and quality needs.

Because of the potentially large data volume and the many-to-many relationship between data providers and potential consumers, performance reasons are clearly an impediment to the conduction of individual data quality assessments for each request made by a potential data consumer.

Moreover, the data provider might not be willing to share his data with a potential data consumer before specific usage conditions are negotiated. On the other hand, the potential data consumer wants to know whether there is a realistic chance that the quality of the data will fit his individual needs prior to deciding whether it is worth starting negotiations with the data provider. Because the quality needs of different data consumers can differ significantly, providing the results of generic quality assessment would not be very useful.

In order to overcome this deadlock, the SQA⁴BD approach decouples the *quality analysis* (i.e., the collection of quality-related measurement results), which depends on the data, from the final individual *quality evaluation*, which depends on the specific quality needs of the potential data consumer.

Fig. 2 depicts an overview of the overall process followed by the SQA⁴BD approach with its key activities for the authority (A1), the data provider (P1 to P5), and the potential data consumer (C1 to C6). If no general authority exists in the considered ecosystem, the data provider may take over the function of the authority by publishing a base quality model.

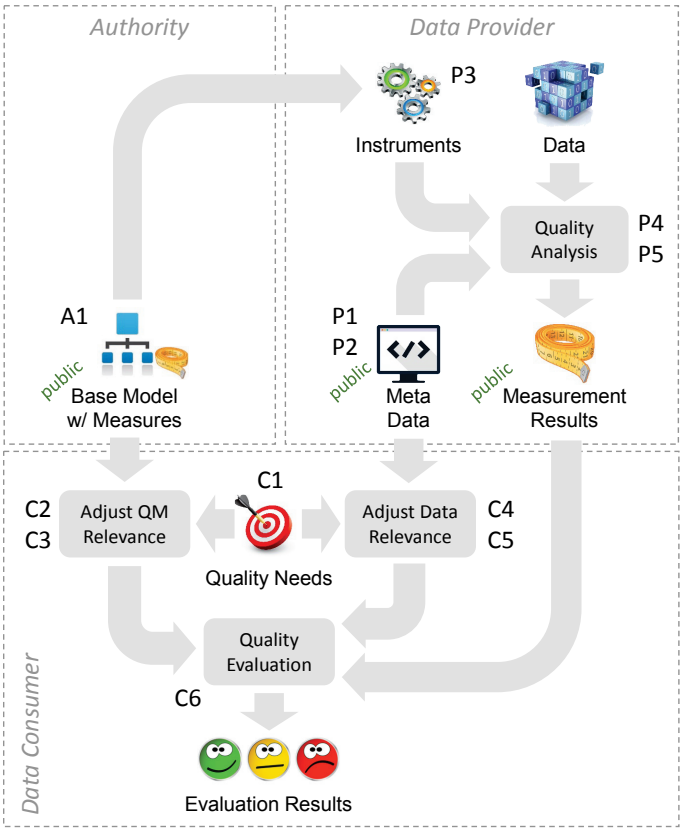

Fig. 2. Conceptual components of a data quality assessment

The idea of adapting a base quality model based on individual quality needs was borrowed from the software quality assessment context, where it was realized and evaluated in the QUAMOCO approach [17][33].

B. Authority: Providing Base Quality Model with Measures

Provide Base Quality Model (A1): The authority decides on a quality model that refines quality with the help of quality aspects (such as completeness, consistency, and timeliness) that are relevant in the considered domain. Quality aspects may be taken from existing standards such as ISO/TS 8000 [14] and ISO/IEC 25012 [13] or one of the various data quality models published in the literature [2]. In order to include a generic level that abstracts from the specific type of data (e.g., database tables, log, xml, or jpg files), quality factors are introduced as a new concept in data quality modeling. The idea is based on the successful use of product factors in the context of software quality evaluation, where they are applied in a similar way to abstract from the specific characteristics of different programming languages. The concept is illustrated in Fig. 3 on an excerpt of a quality model. Depending on the type of data, different measures can be provided to quantify it. For instance, the product factor *absence of attribute values* can be measured by the percentage of records with NULL or dummy values for a given attribute for database tables and by the percentage of records (i.e., individual log messages) without a time stamp.

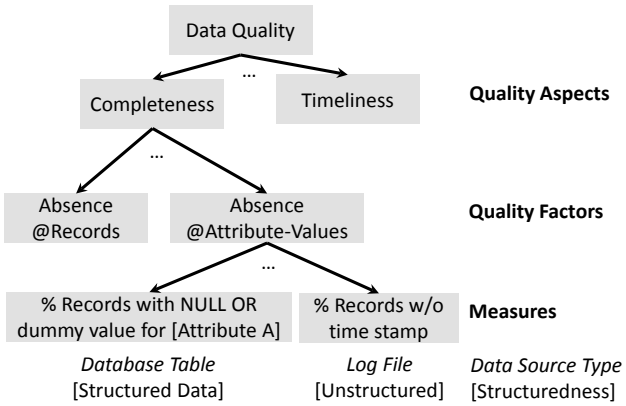

Fig. 3. Excerpt of data quality model structure

C. Provider: Analyzing Data Quality

Provide Meta-Data (P1): In order to make the data usable for others and enable a reasonable quality evaluation, the data provider has to provide data about the data (i.e., *meta-data*). Meta-data comprises:

- *administrative* information, e.g., how the data can be accessed, who is permitted to access the data, whether the data are anonymized or encrypted;
- *structural* information characterizing the internal structure of the data; e.g., in the case of structured data, the table, its attributes and relationships, but also the type of data provided by an attribute (e.g., integer, float, date, text) and consistency rules such as that a specific attribute is not allowed to contain negative values;
- *descriptive* information that allows a deeper understanding of the data, e.g., the organization providing the data, when the data was collected, the meaning of the attributes, and the unit of measurement for numeric data or the specification of applied categories.

Meta-data can have three functions in the context of a quality assessment: (1) It can be required in order to *enable the assessment*, e.g., by defining how the data can be accessed during the assessment; (2) it can consist of *assertions* that are checked during the quality assessment, e.g., the constraint that an attribute does not contain negative values; and (3) it can also consist of *quality-bearing entities*, e.g., by checking that descriptions are available for all relevant attributes.

In our current prototype, relevant meta-data is provided in xml files with a specified schema.

Provide Mapping Information (P2): Because the provided big data is usually heterogeneous and distributed over several information sources, mapping between the data is necessary, for instance to check the data with regard to inconsistencies.

In our current prototype, we provide such mappings together with the meta-data in xml files. An example of such a mapping is the link between the product identifier as used in the log files and the product id column in the database table.

Provide and Configure Instruments (P3): Because the base model and its measures are independent of the concrete infrastructure of the data provider and the format of the data, the measures have to be implemented using specific instruments based on the assessed data. For example, a check for records with NULL or dummy values (cf. Fig. 3) has to implemented differently depending on whether the data is provided in a relational database or in a comma-separated value (csv) file. Moreover, measures have to be implemented for unstructured data such as log files that usually contain heterogeneous types of log messages and do not have to follow a fixed schema. For example, the files could contain log messages about users, performed actions, errors that occurred. Depending on the kind of data, existing ETL tools can help to preprocess data provided in different formats to reduce the number of implementation variants needed for a measure. However, such preprocessing might affect the performance of the assessment compared to a native implementation of the measure.

Moreover, instruments may have to be configured based on the assessed data. For example, it might be necessary to define the data values that are counted as missing values (e.g., NULL, -99, "dummy").

Initial Quality Analysis (P4): During the initial quality analysis, measurement is performed for the considered data sources using the configured instruments defined in P3. The measurement results are stored using the structure of the base model (upper part of Fig. 4) as well as meta-data specifying the data sources and data structure (lower part of Fig. 4). The dark path in Fig. 4 shows an excerpt from an example in which measurement data about missing time stamps is collected for web log messages from two different webservers as data sources.

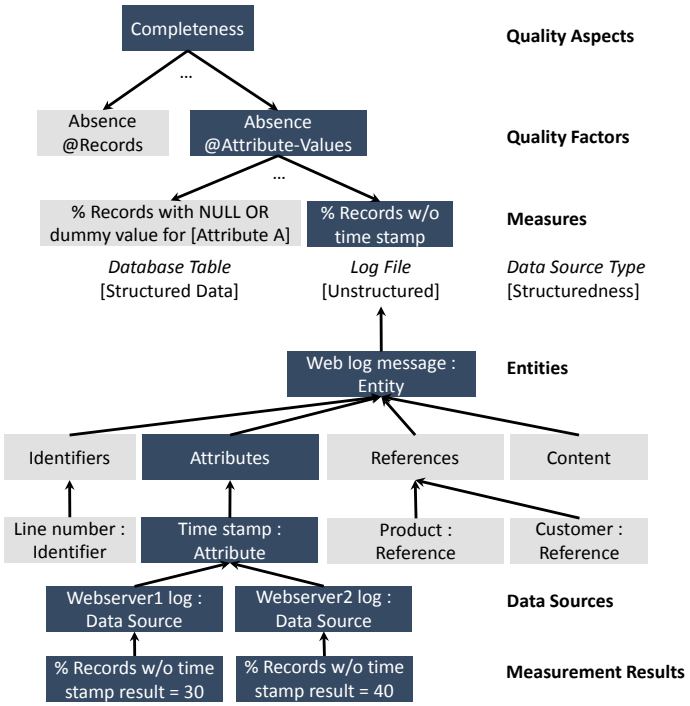

Fig. 4. Excerpt of data structure used to provide measurement results

Updating Quality Analysis (P5): Big data may be updated frequently (cf. velocity), and together with the data, their quality may also change. This means that the quality measurement results need to be updated to remain valid. Due to the data volume, analyzing the complete data each time would be inefficient. Therefore, our approach considers *delta measures*, which analyze only the updated part of the data to recalculate an updated measurement result for the complete data. However, this requires that the measurement definition allows delta calculations, that the updated parts of the data are known, and typically, that some internal calculation results from the initial quality analysis are stored. An interesting observation in this context is that immutable data stores become more common in Big Data environments. The fact that they only allow to add new data but not to modify or delete existing data ease the calculation of delta measures.

Fig. 5 illustrates the approach for the *% records without time stamp* measure from Fig. 3 assuming that existing log messages are neither modified nor deleted: The measure (m_U) is calculated only for the updated part of the data – in this case the additional weblog messages – and then used together with the numbers of previous (n_A) and additional messages (n_U) to update the previous measurement result (m_A). The illustrated delta calculation could also be extended to situations with updated and deleted records but would require in this case a significantly more complex equation. Although we have checked the feasibility to conduct incremental updates for several measures on the conceptual level, we have not yet implemented them in our initial prototype.

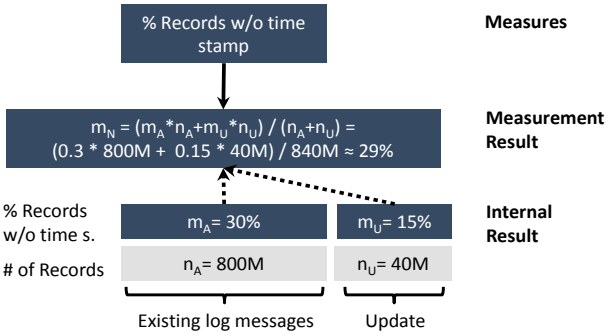

Fig. 5. Example for updating results using data measurement

D. Consumer: Evaluating Data Quality

The quality evaluation approach is based on the general procedure of MCDA (multi-criteria decision analysis) [15] and its specific implementation in the QUAMOCO approach for software quality [33]. Therefore, we provide only a high-level description in this paper, highlighting the differences to the existing approach and refer to [32] regarding technical details.

Identify Quality Needs (C1): The quality needs of a data consumer depend on their information needs, which depends on the purpose and context of the data usage, the user, and the point in time when the data are to be used. Based on these pieces of information, preferences are derived regarding quality aspects and factors as well as the relevance of specific measures, data elements, and sources.

Specify Quality Preferences (C2): Specific quality aspects, factors, and measures may be more relevant than others depending on the identified quality needs. This relevance can be expressed and considered in the quality evaluation results by giving different quality aspects, factors, and measures different weights expressing their relative importance. Weights can be determined directly or calculated based on a ranking provided by the potential user of the data. Quality aspects, factors, and measures identified as irrelevant based on the specific quality needs can be excluded from the evaluation by setting their weights to zero.

Specify Thresholds (C3): Evaluation functions are responsible for translating measurement results, which may be provided on arbitrary measurement scales, into a unified evaluation scale. For the sake of simplicity, we chose the interval [0, 1], where 0 is the worst possible and 1 the best possible evaluation result.

Usually, an evaluation function has to be parameterized by defining thresholds for which the measurement result leads to the worst and best evaluation result, respectively. These thresholds may differ depending on the specific quality needs. For instance, different rates of missing values may be tolerated depending on the application purpose.

Specify the Relevance of Data Elements (C4): Certain data elements such as entities or attributes may have different degrees of relevance depending on the specific quality needs. For instance, the dimensions of a product might be more important than its material if logistics processes are being optimized. Unlike the QUAMOCO approach for software quality, this approach therefore also allows defining specific weights for data elements.

Specify the Relevance of Data Sources (C5): The relative importance of a data source is relevant if data regarding the same entity is obtained from different data sources, for instance if there are multiple databases with customer data. In most cases, the relative relevance of a data source with respect to a specific entity can be determined automatically based on the relative number of records it contains for this entity, but there are situations in which the weights have to be adjusted (e.g., if one data source contains current customers and one contains former customers).

Calculate Quality Evaluation (C6): Based on the measurement results provided by the data provider and the weights and thresholds specified by the potential data consumer (C2 to C4), an individual quality evaluation result can be calculated considering the specific quality needs of the data consumer. Fig. 6 illustrates the calculations on an excerpt. The calculation makes use of the parameterized evaluation functions to map the measurement results on the evaluation scale and aggregates the evaluation results across the evaluation hierarchy based on weighted sums.

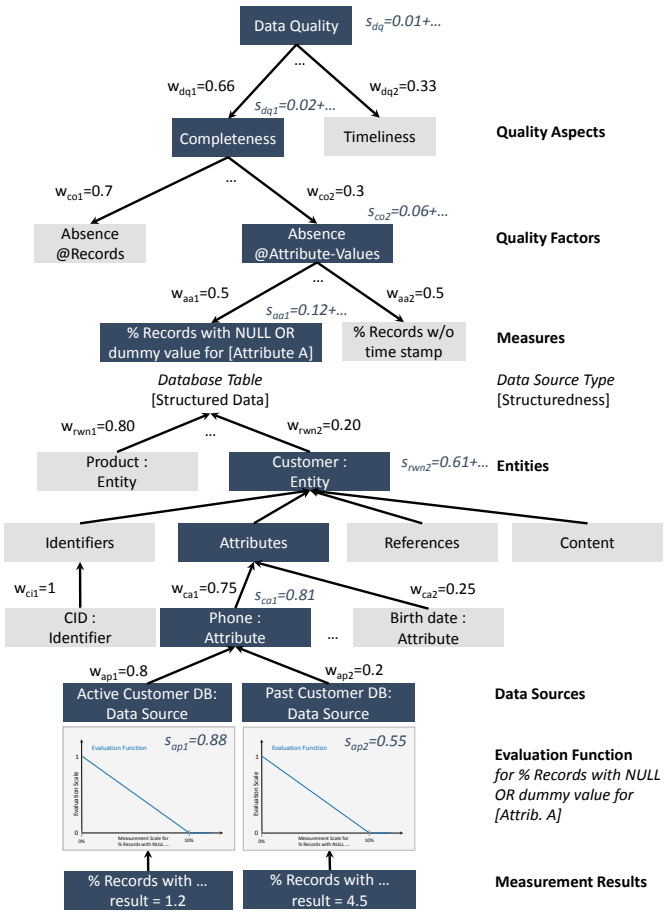

Fig. 6. Example quality evaluation based on measurement results

V. Preliminary Evaluation

As a first evaluation of the developed concepts, a prototype realization of the SQA⁴BD approach was implemented. The focus of the evaluation was the scalability of the approach for large volumes of heterogeneous data. The extent to which the scalability requirement can be realized in the context of the selected system architecture was showcased by implementing a small selection of quality measures.

For the implementation, the Apache Spark Framework with Java, which is also the basis of the PRO-OPT architecture, was used, as well as Apache Hadoop for distributed data management. The implemented quality measures included, in particular: measures for completeness (counting missing data values) and for consistency (checking validity of relationships between records from different data sources).

A sales use case was selected as the application scenario. To get a near to realistic database using a simulator, customer master data, product master data, as well as sales data were generated. In addition, appropriate sales transaction log files were produced. To evaluate the influence of the data volume on the quality assessment performance, we generated test datasets with quality issues of five different sizes (10 GB, 20GB, 30GB, 40GB, and 50GB).

To archive a realistic computation complexity, the calculated consistency measures required to identify contradicting values in disjoint datasets. As part of the calculation, the two respective datasets had to be joined by one or more unique identifier that occurs in both datasets, e.g., a transaction id or a combination of a user id and a timestamp, which has first to be extracted from the analyzed log files.

To evaluate the scalability of the selected realization based on Spark and Hadoop in a distributed environment, a test cluster of five computer nodes was used, each equipped with Intel i5-5250U (2 core) CPU, 16GB memory, 500GB SSD, 1500GB HDD storage, and 1GBit network interface. In the Hadoop file system, the test datasets were provided with a replication factor of three and were analyzed by the quality measures implemented in Java and executed by Spark. In addition, meta-data for the test datasets were provided. Fig. 7 illustrates the overall test setup.

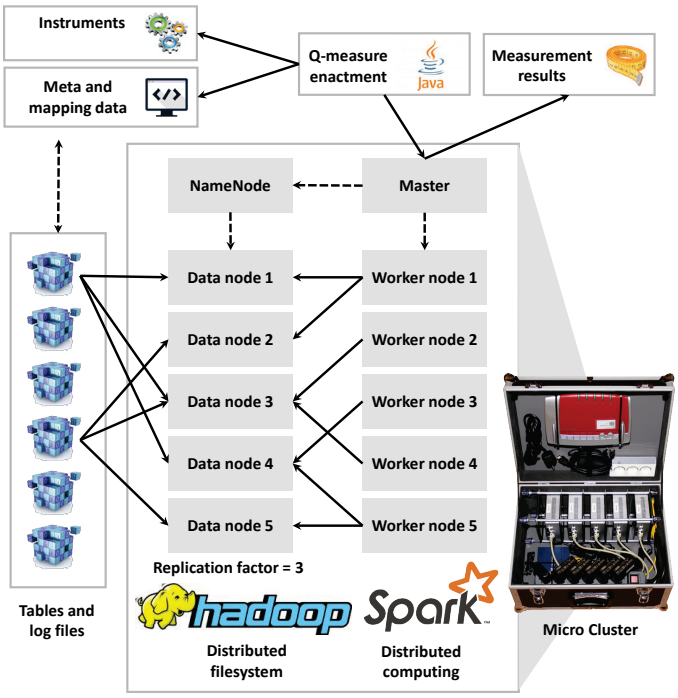

Fig. 7. Test setup of the prototype evaluation

The evaluation comprised weak scalability as well as strong scalability of the realization. *Weak scalability* describes the behavior of the system if the available resources increase to the same extent as the problem size. For that, the runtimes of the combinations of 10GB and 1 Spark node to 50GB and 5 Spark nodes were measured (see Fig. 7). *Strong scalability* describes the savings in terms of runtime when the problem size remains constant and the number of resources increases. Strong scalability was measured at constant problem sizes of 10GB to 50GB while increasing the number of computation nodes from 1 to 5 (see Fig. 7).

The results show that proportionally increasing the problem size (that is, the amount of analyzed data) and the number of computation nodes also increases the runtime of the assessment. However, this runtime increase is not proportional to the

increase in the amount of data, which would result in a decrease of weak scaling efficiency to 20% in the case of five nodes. Instead, a decrease of weak scaling efficiency to 91% can be seen (Fig. 8). Due to the limited technical setup of only five computation nodes and, as a consequence, a limited number of available data points, there is no clear evidence of a functional relationship between the number of computation nodes and weak scalability efficiency (Fig. 9). Therefore, no conclusions can be drawn regarding weak scaling efficiency in the case of a large number of nodes (for example, 100). To this end, further experiments are needed.

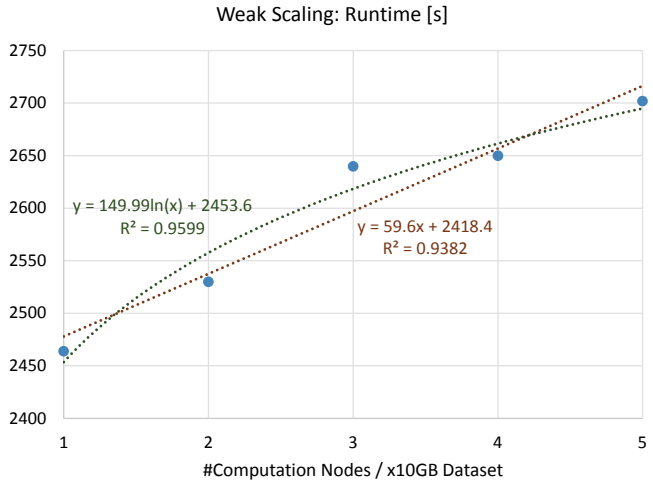

Fig. 8. Runtime in the case of equal increase in data volume and number of computation nodes

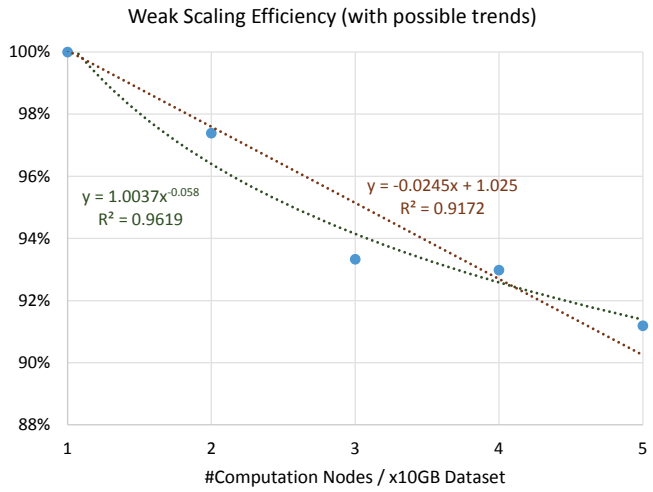

Fig. 9. Gradient of weak scalability efficiency when data volume and number of computation nodes increases

The evaluation of strong scalability efficiency shows a decrease in runtime, which, for a given amount of data, can be very accurately approximated to an exponential function (cf. Fig. 10). Both, the multiplicative factor and the coefficient seems to be dependent on the size of the dataset. The larger the dataset is, the closer the coefficient is to the ideal value of -1. A coefficient of -1 correspond to an anti-proportional decrease of the runtime when adding additional nodes and thus to a strong scaling efficiency of 1.

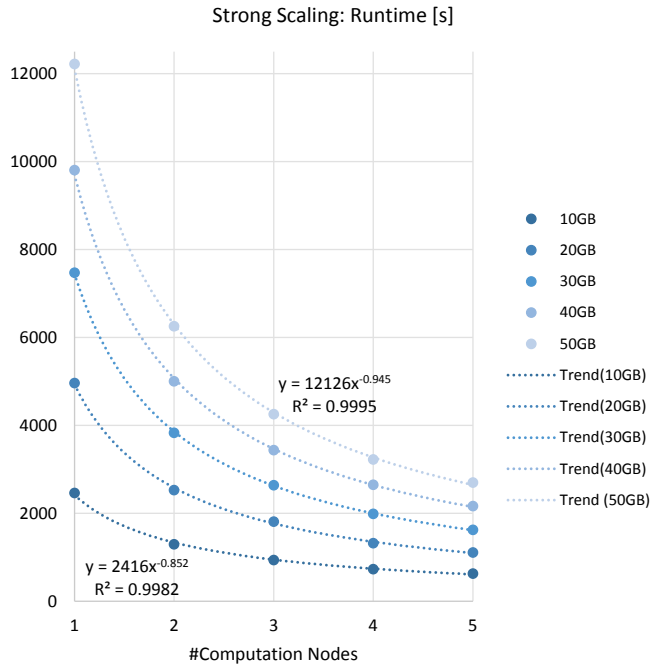

Fig. 10. Runtime when number of computation nodes increases

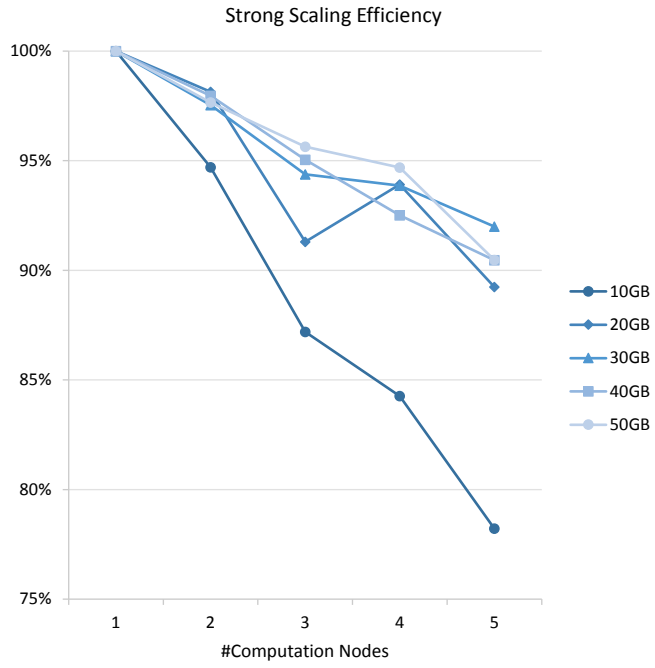

Fig. 11. Change of strong scalability efficiency when number of computation nodes increases

In Fig. 11, which shows the change of strong scalability efficiency when the number of computation nodes increases, especially the runaway values at 30 GB for three and five computation nodes are notable. Compared to Fig. 10, strong scalability efficiency appears sensitive to minor deviations and measurement errors. However, in contrast to weak scalability efficiency, due to the high quality ($R > 0.99$) of the exponential trend models (see Fig. 10) an approximation appears to be possible even for larger numbers of computation nodes. Again, from a statistical point of view it should be noted that each of the models is only based on five data points.

For example, based on the results shown in Fig. 10, the runtime for 50 GB data and 100 computation nodes would result in 156s of runtime, and thus in a strong scalability efficiency of $12126s / (156s * 100) \approx 78\%$.

Further experiments of the scalability behavior, in particular the use of sampling to further reduce the runtime in the case of high volumes of data, can be found in [22].

VI. Conclusion

In this paper, we introduced challenges for data quality assessments caused by the properties of big data (volume, variety, and velocity) in cross-organizational usage scenarios, such as those emerging particularly in smart ecosystems. Because existing approaches do not sufficiently address these challenges, we proposed the new quality assessment approach SQA^{4}BD, which distinguishes between the three roles authority, data provider, and data consumer. Based on these roles, we showed how the approach decouples quality measurement and evaluation to allow efficient quality assessments also for many-to-many relationships between data providers and consumers. Moreover, we illustrated how the approach makes use of delta measurement to allow efficient updating of measurement results in the face of frequently changing data and permits considering various data types in the same model using quality factors as an additional layer of abstraction. Finally, we investigated the horizontal scalability of the approach on an initial prototype using big data technology. The results indicate that although no perfect linear scaling can be observed, increasing volumes of heterogeneous data can be handled by adding additional computation nodes.

The presented version of the approach is limited in the way in which it can deal with changes that do not address the data or quality needs, but the structure of the data or the quality model. However these are important open points that need further investigations in the future. As part of the PRO-OPT project, we plan to further improve the initial prototype by increasing the number of quality measures and improving the usability of the approach. Moreover, we plan to further evaluate the approach on a more powerful cluster infrastructure to validate our first findings and apply it to support industry stakeholders in dealing with their specific big data quality issues.

Acknowledgment

This work is being partially funded by the German Federal Ministry for Economic Affairs and Energy in the context of the technology program "Smart Data - Innovations in Data", grant no. 01MD15004E and by the Ministry of Education and Research in the context of the "Industrial Data Space" project, grand no. 01IS15054.

References

[1] Angeles, P., & MacKinnon, L. M. (2005). Quality Measurement and Assessment Models Including Data Provenance to Grade Data Sources. In *International Conference on Computer Science and Information Systems*.

[2] Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM Computing Surveys*, *41*(3).

[3] Boufares, F., & Ben Salem, A. (2012). Heterogeneous data-integration and data quality: Overview of conflicts. In *6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications*, SETIT 2012.

[4] Buck, D. (2012). Datenqualität, K.o.-Kriterium für Business Intelligence [Data quality, knock-out criteria for business intelligence]. *Computerwoche*. Available: http://www.cowo.de/a/1938325

[5] Cai, L., Zhu, Y. (2015). The challenges of data quality and data quality assessment in the big data era. *Data Science Journal*. 14

[6] Chen, C. C., & Tseng, Y.-D. (2011). Quality evaluation of product reviews using an information quality framework. *Decision Support Systems,* 50(4), 755–768.

[7] Du, J., & Zhou, L. (2012). Improving financial data quality using ontologies. *Decision Support Systems*, 54(1), 76–86.

[8] Even, A., & Shankaranarayanan, G. (2009). Dual assessment of data quality in customer databases. *Journal of Data and Information Quality,* 1(3), 15:1-15:29.

[9] Hubauer, T., Lamparter, S., Roshchin, M., Solomakhina, N., & Watson, S. (2013). Analysis of data quality issues in real-world industrial data. In *PHM 2013 - Proceedings of the Annual Conference of the Prognostics and Health Management Society 2013*.

[10] Ihaka, R., & Gentleman, R. (1996). R: a language for data analysis and graphics. *Journal of computational and graphical statistics*, 5(3), 299-314.

[11] Industrial Data Space (2016). Industrial Data Space – Data Sovereignty. Available: https://www.fraunhofer.de/en/research/lighthouse-projects-fraunhofer-initiatives/industrial-data-space.html

[12] ISO/IEC 25010:2011. Systems and software engineering – Systems and software product Quality Requirements and Evaluation (SQuaRE) – System and software quality models.

[13] ISO/IEC 25012:2008. Software product Quality Requirements and Evaluation (SQuaRE) — Data quality model.

[14] ISO/TS 8000-1:2011. Data Quality – Part1: Overview.

[15] J. Dodgson, M. Spackman, A. Pearman, L. Phillips (2000). Multi-criteria analysis: a manual. Technical report, Department of the Environment, Transport and the Regions, London.

[16] Kan, S. H. (2002). *Metrics and models in software quality engineering*. Addison-Wesley Longman Publishing Co., Inc.

[17] Kläs, M., Lampasona, C., Münch, J. (2011). Adapting software quality models: practical challenges, approach, and first empirical results. In *37th EUROMICRO Conference on Software Engineering and Advanced Applications* (SEAA), Oulu, Finland, IEEE Computer Society.

[18] Kläs, M., Putz, W., Lutz, T. (2015) Überblick und Potentialanalyse bestehender Modellierungsansätze zur Datenqualität [Overview and Potential Analysis of existing modelling approaches for data quality], IESE Report.

[19] Kläs, M., Trendowicz, A., Jedlitschka, A. (2015). What Makes Big Data Different from a Data Quality Assessment Perspective? Practical Challenges for Data and Information Quality Research. IESE-Report 071.15/E.

[20] Laney, D. (2001). 3D data management: controlling data volume, velocity, and variety. Available: http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

[21] LimeSurvey (2015). LimeSurvey - the most popular free open source software survey tool on the web. Available: https://www.limesurvey.org

[22] Lutz, T. (2015). A quality assessment approach for large heterogeneous data and its operationalization. Master Thesis, Technical University Kaiserslautern.

[23] Markl, V., Hoeren, T., Krcmar, H. (2013). Innovationspotenzialanalyse für die neuen Technologien für das Verwalten und Analysieren von großen Datenmengen (Big Data Management).

[24] Martin, N., Poulovassilis, A., & Wang, J. (2014). A Methodology and architecture embedding quality assessment in data integration. *Journal of Data and Information Quality,* 4(4), 17:1-17:40.

[25] MySQL (2015). MySQL – The world's most popular open source database. Available: http://www.mysql.com

[26] Petersen, K., Feldt, R., Mujtaba, S., Mattsson, M. (2008). Systematic mapping studies in software engineering. In *International Conference on Evaluation and Assessment in Software Engineering*

[27] PRO-OPT (2016). Big Data Produktionsoptimierung in Smart Ecosystems. Available: http://www.pro-opt.org

[28] Reznik, L., & Bertino, E. (2013). Poster: Data quality evaluation: Integrating security and accuracy. In *Proceedings of the ACM Conference on Computer and Communications Security*.

[29] Roger Clarke, R. (2014). Quality Factors in Big Data and Big Data Analytics. Available: http://www.rogerclarke.com/EC/BDQF.html

[30] Talend. *Talend Open Studio for Data Quality*. Available: http://www.talend.dreamhosters.com/top/user-guide-download/V562/TalendOpenStudio_DQ_UG_5.6.2_EN.pdf

[31] UNECE Big Data Quality Task Team (2014). A suggested framework for the quality of big data. UNECE Report.

[32] Wagner, S., Goeb, A., Heinemann, L., Kläs, M., Lampasona, C., Lochmann, K., Mayr, A., Plösch, R., Seidlf, A., Streit, J., Trendowicz, A. (2015). Operationalised product quality models and assessment: The Quamoco approach. *Information and Software Technology*, 62, 101-123.

[33] Wagner, S., Lochmann, K., Heinemann, L., Kläs, M. et al. (2012). The Quamoco product quality modelling and assessment approach. In *International Conference on Software Engineering* (ICSE), Zürich.